Classification of phylogenetic data via Bayesian mixture modelling

Loza Reyes, Elisa

Award date:
2010

Awarding institution:
University of Bath

Classification of phylogenetic data via Bayesian mixture modelling

submitted by

Elisa Loza Reyes

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

January 2010

Summary

Conventional probabilistic models for phylogenetic inference assume that an evolutionary tree, and a single set of branch lengths and stochastic process of DNA evolution are sufficient to characterise the generating process across an entire DNA alignment. Unfortunately such a simplistic, *homogeneous* formulation may be a poor description of reality when the data arise from *heterogeneous* processes. A well-known example is when sites evolve at heterogeneous rates. This thesis is a contribution to the modelling and understanding of heterogeneity in phylogenetic data. We propose a method for the classification of DNA sites based on Bayesian mixture modelling. Our method not only accounts for heterogeneous data but also identifies the underlying classes and enables their interpretation. We also introduce novel MCMC methodology with the same, or greater, estimation performance than existing algorithms but with lower computational cost. We find that our mixture model can successfully detect evolutionary heterogeneity and demonstrate its direct relevance by applying it to real DNA data. One of these applications is the analysis of sixteen strains of one of the bacterial species that cause Lyme disease. Results from that analysis have helped understanding the evolutionary paths of these bacterial strains and, therefore, the dynamics of the spread of Lyme disease. Our method is discussed in the context of DNA but it may be extended to other types of molecular data. Moreover, the classification scheme that we propose is evidence of the breadth of application of mixture modelling and a step forwards in the search for more realistic models of the processes that underlie phylogenetic data.

Acknowledgements

I would like to thank my supervisors Tony Robinson and Merrilee Hurn for their help, guidance and support during this time. I will always appreciate our stimulating discussions during our weekly meetings and their careful and constructive criticism of this thesis. I would also like to thank Tony for introducing me, back in 2005, to the fascinating field of statistical phylogenetics, and to both of them for their interest in learning with me some of the intricacies of molecular biology. I am grateful to Gabi Margos, Klaus Kurtenbach and Ed Feil, from the Biology and Biochemistry Department at the University of Bath, for their contagious appetite for understanding evolutionary phenomena, for providing me with large quantities of molecular data to analyse and for their always prompt advice. In addition I would like to thank Tom Nye for his thoughts on a first draft of the introduction to this thesis and Leo Lara for his help with C programming. I am indebted to the Mexican Council for Science and Technology, CONACYT, for its financial support throughout my postgraduate studies, without which this work would not have been possible.

Thanks also to Steve for his love, support and invaluable help in ensuring that some parts of this thesis are comprehensible to non-statisticians. I am also grateful to my family for always being there for me and for being an example of determination and hard work. Finally I would like to thank all my friends and staff from the Department of Mathematical Sciences for making these years in Bath such enjoyable and memorable ones.

Contents

Chapter 1

Introduction

1.1 The problem

Statistical phylogenetics is the inference and interpretation of evolutionary trees and other parameters in order to describe and understand the evolution of a group of organisms. The type of data on which phylogenetic inferences in this thesis are based are the DNA sequences of the organisms. In statistical inference, we attempt to formulate a probabilistic model for the processes that generate these data. This model is usually expressed in terms of three parameters: an evolutionary tree that represents the relationships that hold between the organisms, the lengths of the branches of the tree which symbolise the amount of evolutionary divergence, and a stochastic process that models how a DNA character is substituted by another (e.g. [23]). Recent phylogenetic developments advocate the use of networks in situations where the underlying evolutionary process is complex (e.g. [49]), but in this thesis we will concentrate on the inference of trees.

The observed DNA sequences, each N characters long, are usually 'aligned' on top of each other to form a matrix, or *DNA alignment*. This alignment has as many rows as DNA sequences are observed and as many columns as characters in each sequence. Each column in the alignment is referred to as a *site*. Conventional probabilistic models for phylogenetic inference assume that a single tree, set of branch lengths and stochastic model are sufficient to characterise the evolutionary process across all sites in a DNA alignment. But such a simplistic, *homogeneous* formulation may be inadequate in some cases. It is not uncommon to find DNA alignments in which some sites correspond to one gene and the rest to a different gene, and it is well known that genes evolve at different rates depending on their position in an organism's genome (e.g. [29, 60]). In a case like this, a homogeneous model that supposes that all sites arise from a common evolutionary process is a poor description of reality. An example of the potentially deficient fit of a homogeneous model is found in DNA data arising from both *pseudogenes* and *genes*. Pseudogenes, 'dead' copies of genes that have acquired mutations and gradually ceased functioning, have a different way of evolving to their functional counterparts, the genes. Because pseudogenes do not have

apparent importance in the healthy-functioning of an organism they are freer to change from generation to generation than genes. Whenever *heterogeneous* phylogenetic data are described with a *homogeneous* probabilistic model, inferences are a compromise between the differing evolutionary processes underlying the data.

Consider a dataset containing the DNA sequences for six organisms whose evolutionary history we wish to infer. Suppose that half of the sites arise from a process that obeys the tree at the top-left of Figure 1-1 (class A) and half of them from an evolutionary process that follows the top-right tree in the same figure (class B). The branches of these trees have very different lengths, which suggests that the two processes have a rather distinctive nature. If we interpret branch lengths as the amount of evolutionary divergence between the organisms and think in terms of the sum of all branch lengths, then organisms in class A are much more divergent (on average) from each other than organisms in class B. This means that process A is evolving faster than process B to produce more divergent individuals. This could be due, for example, to class A corresponding to observations that originate from pseudogenes (which are rapidly changing from generation to generation) and class B conforming to functional genes that accumulate changes at a slow rate. By fitting a homogeneous model to these synthetically-produced data the estimated branch lengths, obtained as the sample average of $15\,000$ MCMC iterations after burn-in, are the ones shown at the bottom of Figure 1-1. These lengths are a clear compromise between the two evolutionary processes and they do not reflect the true nature of either class. Biological conclusions derived from these estimates will thus be incorrect.

One typical approach to alleviate the poor fit of a homogeneous model to heterogeneous data is to partition the DNA alignment into classes prior to analysis so that different probabilistic models are assigned to each individual partition. This approach relies on prior knowledge about the class-structure that underlies the data and is far from ideal since such *a priori* information is not always available. An alternative approach is to construct a model that includes one parameter per site so that, for example, each site has its own rate parameter (e.g. the gamma model popularised by Yang [125]; Section 5.2.1). None of these methods, however, provide a framework for partitioning the sites into mutually exclusive classes as part of the same inferential procedure.

The main problem to be addressed in this thesis is the classification of DNA sites into distinct evolutionary classes within a phylogenetic context. The number of classes and the characteristics of the classes need to be determined. Thus a scientist interested in classifying the sites by the tree shape that relates the organisms must characterise the classes in one way, whereas a scientist concerned with the separation of the sites by the branch lengths will require a different delineation of the classes. In a classification scheme of the former type, all sites considered in the example from Figure 1-1 would belong to the same class (they all share the same tree shape), whereas a classification method by the length

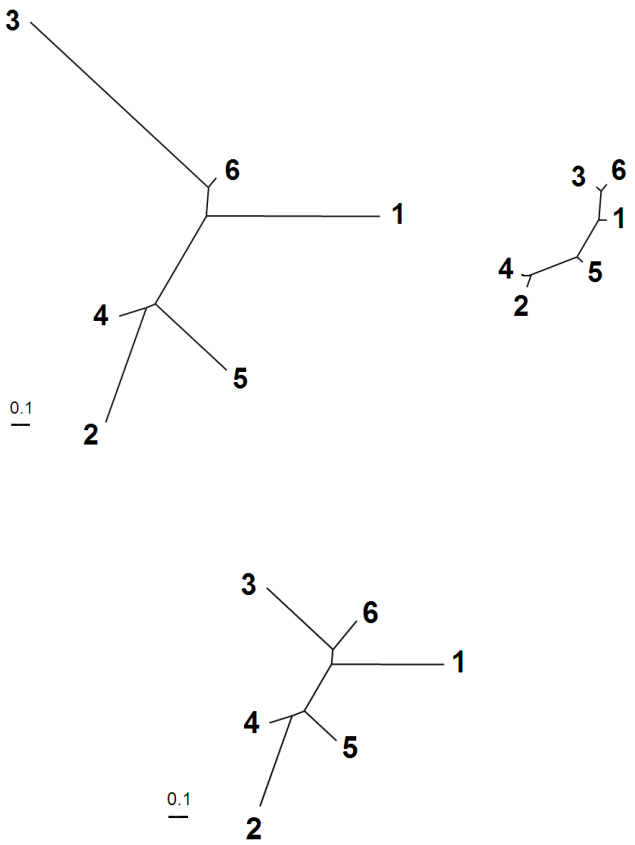


Figure 1-1: The two figures at the top represent the true trees and branch lengths from which DNA data arise. Suppose that half of the sites in the alignment originate from one tree and half from the other. If these data are fitted by a homogeneous model, estimates are a compromise between conflicting evolutionary processes, as shown by the estimated branches of the tree at the bottom. The branch lengths in this tree were estimated as the sample average of an MCMC simulation of $15\,000$ iterations after burn-in.

of the branches would separate the sites into two different groups. Meaningful inferences thus require that the purpose of the classification is clearly specified.

In phylogenetics, the development of models that account for heterogeneity among observables has long been a topic of interest (e.g. [125, 28, 111, 88]). A classification method for phylogenetic data such as the one we propose not only accounts for heterogeneity but also provides us with some insight into the nature of the heterogeneity itself. For example, the classification could be used as an indicator of evolutionary agreement or disagreement between different genes. If two sites known to belong to different genes are classified to the same class, this is evidence of evolutionary compatibility between the genes. On the other hand, if the sites are classified to different categories, then the genes can be suspected of following different evolutionary processes. In scenarios in which the membership of sites to different genes is not *a priori* known, such a classification scheme could still be used to detect evolutionary heterogeneity in the data. A method for the classification of DNA sites is not only convenient as a means of accounting for data generated under *heterogeneous*

conditions but also as a way of identifying the classes and interpreting their nature.

1.2 The aims of the thesis

Finite mixture distributions provide a natural means for modelling data generated under heterogeneous conditions. Their flexibility has encouraged their use in a wide range of applications, including the analysis of galaxy data by Roeder [101]; the modelling of thickness of Mexican stamps by Izenman and Sommer [55]; and the study of data from the Old Faithful geyser in Yellowstone National Park by Silverman [107].

Most of this thesis is dedicated to reformulating the problem of the classification of DNA sites as a process of estimating the parameters of a Bayesian mixture of 'phylogenetic' distributions. The motivation for studying the problem of DNA classification via finite mixture models comes from a publication by Pagel and Meade [88]. They devised a mixture model to account for heterogeneity in phylogenetic data but did not attempt classification as part of a single inferential procedure. The component distributions of our novel mixture model are assumed to conform to the conventional probabilistic model for phylogenetic inference, as formulated by Felsenstein [23]. Each component corresponds to one evolutionary class and the purpose of the analysis is to identify the component from which each site arises. Specifically, we let the branch lengths and the stochastic process of DNA substitution vary between mixture components, and assume a common-to-all-component tree.

This thesis also aims to contribute MCMC methodology for the efficient estimation of phylogenetic parameters. Some phylogenetic parameters pose challenges to MCMC simulation. For example, a tree is not a real-valued parameter but rather a graphical object that must be estimated from the data. The design and implementation of MCMC samplers requires us to find efficient mechanisms for updating the tree, storing it in a computer-readable format and summarising the stream of trees sampled during MCMC simulation into, perhaps, a single 'most likely' tree. The research underlying this thesis has addressed these problems.

Finally, this research aims to validate our novel mixture model with real DNA data, identify the evolutionary classes and, if possible, give a biological interpretation of the classes.

1.3 The structure of the thesis

The field of Bayesian phylogenetic inference is built upon a collection of statistical tools: from statistical modelling itself to the nearly-indispensable simulation via MCMC; from an understanding of real-life phenomena in molecular biology to intensive computer programming. Chapter 2 presents an overview of these tools, all of which are used throughout

this thesis. In Chapter 3, we give a detailed discussion of the conventional, homogeneous model for likelihood-based phylogenetic inference, as formulated by Felsenstein [23]. This homogeneous formulation is the basic component of our mixture model.

In Bayesian inference, we are often interested in expectations of the form

$$\mathbb{E}(f(\theta)) = \frac{\int f(\theta)\, p(\theta) p(x|\theta) d\theta}{\int p(\theta) p(x|\theta) d\theta}$$

where x denotes the observed data, θ denotes model parameters, $p(\theta)$ is a *prior* distribution representing beliefs about the value of θ before any data are observed, and $p(x|\theta)$ is the likelihood for θ. Evaluating this expectation can pose a challenge to statistical inference if the integration is over high-dimensional parameter spaces. Chapter 4 discusses MCMC simulation, an approach to approximate the typically complex expectations that arise in phylogenetic inference. In particular, the chapter reviews some existing MCMC methods for estimating phylogenetic models. We propose that the use of sophisticated and computationally expensive MCMC schemes (such as tempered MCMC) may not be justifiable whenever cheaper methods that achieve the same efficiency are available. We introduce and implement a number of alternative algorithms that achieve the same, or greater, efficiency than existing methods but with lower computational cost. We also consider commonly-used MCMC methods that update tree and branch lengths *en bloc* (or simultaneously), and argue that this may have detrimental effects on the estimation performance of the sampler. Chapter 5 introduces the novel phylogenetic mixture model for classification of DNA sites. Chapter 6 presents the algorithmic aspects of the MCMC sampler employed for estimating this mixture model. The results of two applications of our classification method to DNA data are presented in Chapters 7 and 8. The first of these corresponds to the analysis of the *primate mitochondrial DNA* alignment. This dataset has been extensively used to validate proposed phylogenetic methodologies in the past (e.g. [125, 131, 63, 112]) and is relatively well understood within the phylogenetics community. We show that our model is able to detect evolutionary heterogeneity in this alignment and we also give an interpretation of this heterogeneity. Chapter 8 then introduces a new set of phylogenetic data, the DNA sequences from sixteen different strains of one of the bacterial species that cause Lyme disease. We demonstrate consistency in the evolution of eight genes from the genome of these strains and evolutionary inconsistency between those eight genes and a ninth gene. The results of this analysis have been of direct relevance to scientists at the University of Bath studying the dynamics of the spread of Lyme disease and identifying and monitoring the origins, directions and diversity of these bacterial strains.

Finally, Chapter 9 presents conclusions on the main results from this thesis and assesses their contribution to the field. The strengths and limitations of this work are discussed, with possible extensions and ideas for future work.

Chapter 2

Statistical tools

2.1 Introduction

The purpose of this chapter is to outline the statistical theory used in this research. The increasing complexity in both data structures and phylogenetic models have encouraged more and more a Bayesian orientation to statistical phylogenetics, and this is the statistical framework adopted in this thesis. Posterior distributions arising in Bayesian phylogenetic analyses cannot usually be evaluated analytically and they must be approximated by simulation. MCMC techniques are the most popular choice and they are used extensively for the applications presented in this thesis. In this chapter we review both Markov chain concepts related to MCMC simulation and MCMC methods themselves. We also present a general outline of finite mixture models, whose application for classification of phylogenetic data is one of the main contributions of this research. The remainder of this chapter discusses the nature of data encountered in the analyses underlying this research, presents the reader with general terminology used throughout and finally, comments on the implementation of a computer program for phylogenetic MCMC simulation.

This chapter does not intend to be exhaustive but rather to reflect the breadth of Bayesian phylogenetic inference. There are many excellent sources that provide a more comprehensive treatment of the concepts outlined here. For theory on Markov chains in both discrete and continuous time, see Grimmett and Stirzaker [44]. Key references for MCMC simulation are Hastings [46], Gilks, Richardson and Spiegelhalter [38], Gamerman and Lopes [31], and Chib and Greenberg [14]. Finally, a thorough discussion of Bayesian finite mixture modelling, both before and after the advent of MCMC, can be found in the works by Titterington, Smith and Makov [116], McLachlan and Peel [76], and Robert [96].

2.2 Bayesian statistical inference

In a Bayesian approach to statistical inference, both parameters θ and observed data x are treated as random quantities. The inferential procedure requires us to specify a full

probability model over all random quantities in the form of a joint probability distribution,

$$p(\theta, x) = p(\theta)p(x|\theta), \tag{2.1}$$

where θ is usually a high-dimensional parameter vector with *prior distribution* $p(\theta)$ and $p(x|\theta)$ is the *likelihood function* for θ (interchangeably denoted by $p(x|\theta)$ or $L(\theta|x)$ in this thesis). The prior distribution represents beliefs about the value of θ before any data are observed. These beliefs may be prior ignorance or be based on scientific experience from previous studies of similar data. A prior distribution is typically specified by a standard probability distribution that depends on one or more *hyperparameters* that may or may not be known. Most of the prior distributions employed in our analyses are *noninformative*, which means that the influence of the prior information in the inferential procedure is minimised. All of our prior distributions are *proper*, which in general terms refers to a function $p(\theta)$ that integrates to 1 (or to any positive finite value such that $p(\theta)$ can be multiplied by a constant to integrate to 1) [32].

A Bayesian approach offers the possibility of including prior beliefs about the value of θ. Most importantly, it results in a probability distribution for θ, called the *posterior distribution*, which represents probability statements about θ conditional on the observed data. The posterior distribution is given by,

$$\pi(\theta|x) \;=\; \frac{p(\theta, x)}{p(x)} \;=\; \frac{p(\theta)p(x|\theta)}{p(x)}, \tag{2.2}$$

which is obtained by using Bayes theorem. The denominator $p(x)$ is the marginal distribution of the data and is calculated by summing over all possible values of the parameter, $p(x) = \sum_{\theta} p(\theta, x)$, if θ is discrete, or by integrating it out, $p(x) = \int_{\theta} p(\theta, x)d\theta$, if θ is continuous. Often, we are interested in the posterior expectation

$$\mathbb{E}_{\pi}(f(\theta)|x) = \int_{\theta} f(\theta)\pi(\theta|x)d\theta \tag{2.3}$$

of a function f of θ (with the integral substituted by a summation for discrete θ). In very simple cases, expectation (2.3) can be evaluated explicitly but in complex applications its evaluation may involve high-dimensional integrals. In phylogenetic analyses, for example, the region of integration is usually a vast, multi-dimensional space which makes the explicit calculation of (2.3) mathematically intractable. It may also occur that the posterior distribution is only known up to proportionality, as the denominator in distribution (2.2) may be unknown. Expectation $\mathbb{E}_{\pi}(f(\theta)|x)$ can then be approximated by drawing a large number of samples $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}\}$ from π and letting

$$\mathbb{E}_{\pi}(f(\theta)|x) \approx \frac{1}{M}\sum_{n=1}^{M} f(\theta^{(n)}). \tag{2.4}$$

Sampling from π can be achieved in a number of different ways, the simplest of which are direct methods such as sampling by inversion or by rejection (see for example [95]). If direct sampling from π is unfeasible, an alternative technique is *importance sampling*. In importance sampling, an over-dispersed distribution φ with the same support as π but from which it is possible to sample directly, is used instead. Direct and importance sampling are preferred over other alternatives because they provide a convenient and computationally efficient way of generating independent samples. The majority of the problems in phylogenetics though, are too complex for these methods to be of any use and a different approach is needed.

MCMC simulation provides the means to evaluate expressions of the form (2.3) in situations where π is quite non-standard. The idea is to set up a Markov chain that has π as its unique stationary distribution, then to simulate from the Markov chain and, as the chain progresses, we expect to generate samples from a distribution that becomes closer and closer to π. The sample output can then be used in (2.4) to approximate the desired expectation. Since π is the distribution we wish to sample from, we refer to it as the *target distribution*.

In the next section, an overview of discrete-time Markov chains is presented as a preliminary step to discussing MCMC simulation techniques.

2.3 Discrete-time Markov chains

Let $\{\theta^{(n)} : n \in T\}$ be a collection of random variables where T is an *index set* that represents the iterations of a simulation scheme and the variable $\theta^{(n)}$ takes values in a set Θ, called the *state space*, for all $n \in T$. Throughout, the index set is taken to be the set of natural numbers and the state space is assumed to be a discrete set. Then, it is said that $\{\theta^{(n)} : n \in T\}$ is a discrete-time *stochastic process* with discrete state-space Θ.

A *Markov chain* is a stochastic process with the property that the future behaviour of the process is independent of the past and only depends on the present state. That is, for all integers $n \in T$ and all states $i_0, \ldots, i_{n-1}, i, j \in \Theta$,

$$Pr(\theta^{(n+1)} = j | \theta^{(0)} = i_0, \ldots, \theta^{(n-1)} = i_{n-1}, \theta^{(n)} = i) = Pr(\theta^{(n+1)} = j | \theta^{(n)} = i). \qquad (2.5)$$

It is convenient to concentrate on processes where the probabilities in (2.5) are time invariant, that is, situations in which the right-hand side in (2.5) is not dependent on n but only upon i and j. Then, $Pr(\theta^{(n+1)} = j | \theta^{(n)} = i) = Pr(\theta^{(1)} = j | \theta^{(0)} = i)$ and this process is called a *homogeneous Markov chain*. Throughout, all Markov chains are assumed to be homogeneous.

The *one-step transition probability*, $p(i,j)$, indicates the probability that the process is in state j at step one given that the process was in state i at the previous time step. It is defined as

$$p(i,j) = Pr\big(\theta^{(1)} = j | \theta^{(0)} = i\big). \tag{2.6}$$

This probability describes the evolution of the chain at consecutive time steps and satisfies

$$p(i,j) \geq 0 \qquad \text{and} \qquad \sum_{k \in \Theta} p(i,k) = 1 \tag{2.7}$$

for all $i, j \in \Theta$. The *n-step transition probability*, denoted by $p^n(i,j)$, is the probability that the chain moves from state i to j in exactly n time steps. It is defined as

$$p^n(i,j) = Pr(\theta^{(n)} = j | \theta^{(0)} = i), \tag{2.8}$$

and for completeness, $p^1(i,j) = p(i,j)$ and $p^0(i,j) = I[i=j]$, where $I[\cdot]$ is the indicator function taking value 1 when its argument is true and 0 otherwise.

Markov chains for MCMC simulation are required to be irreducible, aperiodic and positive recurrent. These properties are explained in the following sections.

2.3.1 Irreducibility

State j can be reached from state i if there is a non-zero probability $p^n(i,j) > 0$, for some $n > 0$. In that case, we write $i \to j$. If i can be reached from j ($j \to i$) and if j can be reached from i ($i \to j$), it is said that i communicates with j or that i and j are in the same communication class ($i, j \in \Theta$). We then write $i \leftrightarrow j$, which implies that there exist integers $n_1, n_2 > 0$ such that $p^{n_1}(i,j) > 0$ and $p^{n_2}(j,i) > 0$.

The class containing state i consists of all those states $j \in \Theta$ such that $i \leftrightarrow j$. If all states are in the same class, that is, if every state can be reached from every other state in some finite number of time steps, then there is only one communication class. If there is only one communication class the Markov chain is said to be *irreducible*.

2.3.2 Aperiodicity

In order to illustrate the property of aperiodicity, we first introduce a useful object. The *transition graph* of a Markov chain is the graphical structure that represents the relationships that hold between states. To each state there corresponds a node of the graph. A directed edge from node i to node j indicates that state j can be reached from state i or, equivalently, $p(i,j) > 0$.

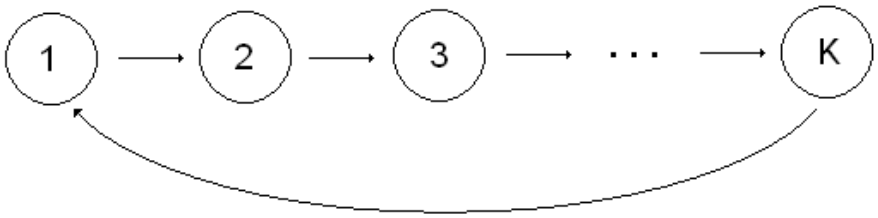


Figure 2-1: The transition graph of a Markov chain with state space $\Theta = \{1, 2, \ldots, K\}$. Beginning at state $i \in \Theta$, it takes exactly K steps to return to state i and we say that the chain has period K.

The transition graph of a Markov chain with state space $\Theta = \{1, 2, \ldots, K\}$ is shown in Figure 2-1. This chain has the particularity that beginning in state i, it takes exactly K steps to return to state i. We say that the chain is *periodic* with period K because it moves across the state space in a regular, cyclic manner. More formally, for some $i \in \Theta$, the period of an irreducible Markov chain is the greatest common divisor of all integers $n > 0$ such that $p^n(i, i) > 0$. If the period is greater than one, the chain is said to be *periodic*. If the period equals one, then the chain is said to be *aperiodic*.

2.3.3 Positive recurrence

Finally, positive recurrence means that the expected time of the first return to state $i \in \Theta$ is finite. A condition for positive recurrence is the existence of a probability distribution $\nu = \{\nu(j) : \ j \in \Theta\}$ that satisfies

$$\sum_{i \in \Theta} \nu(i) p(i, j) = \nu(j), \quad \text{for all } j \in \Theta \qquad \text{(global balance)}.$$

Here ν is called a *stationary probability distribution* of the Markov chain with transition probabilities $p(i, j)$, for $i, j \in \Theta$. As global balance is hard to check, we usually verify instead the condition of 'detailed balance' (also known as 'time-reversibility'):

$$\nu(i) p(i, j) = \nu(j) p(j, i), \quad \text{for all } i, j \in \Theta \qquad \text{(detailed balance)}$$

which implies general balance by taking the sum over i on both of its sides. Construction of a Markov chain with a prescribed stationary distribution π (the target distribution of interest) reduces to finding a law of evolution $p(i, j)$ that satisfies detailed balance with respect to π. In the next section, two commonly used algorithms for the construction of such laws are presented. If in addition the chain is irreducible and aperiodic, then the stationary distribution is unique and the chain converges to it in the long run (see for example [44, ch. 6]).

2.4 MCMC simulation

MCMC simulation provides the means whereby to sample from a complex target distribution. MCMC can be implemented via the Metropolis-Hastings algorithm [78, 46] or the Gibbs sampler [33]. These algorithms allow for a framework in which a typically highly multidimensional parameter vector, $\boldsymbol{\theta}$, is partitioned into d distinct components and only one component is updated at a time. A single *iteration* consists of d sequential updating *steps*, which may be all Metropolis-Hastings, all Gibbs, or some of one sort and the rest of the other. In the following paragraphs, we briefly introduce both the Metropolis-Hastings algorithm and the Gibbs sampler. However, because the Gibbs sampler is a particular case of a Metropolis-Hastings algorithm, most of our discussion will be related to the latter.

2.4.1 The Metropolis-Hastings algorithm

Algorithm 2.4.1. *Given a starting point $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_h^{(0)}, \ldots, \theta_d^{(0)})$, for each iteration $n = 1, 2 \ldots, M$, carry out steps 1-3 for each component in $\boldsymbol{\theta}$. For component h:*

1. *Set $\theta_h = \theta_h^{(n-1)}$.*
2. *Generate a candidate point θ_h' according to an arbitrary proposal distribution $q(\theta_h, \theta_h')$.*
3. *Accept θ_h', that is set $\theta_h^{(n)} = \theta_h'$, with probability*

$$\alpha(\theta_h, \theta_h') = min\left(1, \frac{\pi(\theta_h'|\theta_{-h}^{(n-1)}, x)}{\pi(\theta_h|\theta_{-h}^{(n-1)}, x)} \frac{q(\theta_h', \theta_h)}{q(\theta_h, \theta_h')}\right) \tag{2.9}$$

where

$$\theta_{-h}^{(n-1)} = (\theta_1^{(n)}, \ldots, \theta_{h-1}^{(n)}, \theta_{h+1}^{(n-1)}, \ldots, \theta_d^{(n-1)}) \tag{2.10}$$

represents all the components of $\boldsymbol{\theta}$, except for θ_h, at their current values. So, for the components already updated, their latest value is at iteration n while for all other components is at iteration $n-1$.

Reject θ_h', that is set $\theta_h^{(n)} = \theta_h$, with probability $1 - \alpha(\theta_h, \theta_h')$.

The Metropolis-Hastings algorithm defines transition probabilities of the form

$$p(\theta_h, \theta_h') = \begin{cases} q(\theta_h, \theta_h')\, \alpha(\theta_h, \theta_h'), & \theta_h' \neq \theta_h \\ q(\theta_h, \theta_h) + \sum\limits_{\theta_h'' \neq \theta_h} q(\theta_h, \theta_h'')[1 - \alpha(\theta_h, \theta_h'')], & \theta_h' = \theta_h \end{cases}$$

Detailed balance with respect to π can be easily verified by

$$\begin{aligned}
\pi(\theta_h|\cdot)\, p(\theta_h, \theta_h') &= \pi(\theta_h|\cdot)\, q(\theta_h, \theta_h')\, min\left(1, \frac{\pi(\theta_h'|\cdot)}{\pi(\theta_h|\cdot)} \frac{q(\theta_h', \theta_h)}{q(\theta_h, \theta_h')}\right) \\
&= min\big(\pi(\theta_h|\cdot)\, q(\theta_h, \theta_h'),\, \pi(\theta_h'|\cdot)\, q(\theta_h', \theta_h)\big) \\
&= \pi(\theta_h'|\cdot)\, q(\theta_h', \theta_h)\, min\left(1, \frac{\pi(\theta_h|\cdot)}{\pi(\theta_h'|\cdot)} \frac{q(\theta_h, \theta_h')}{q(\theta_h', \theta_h)}\right) \\
&= \pi(\theta_h'|\cdot)\, p(\theta_h', \theta_h)
\end{aligned}$$

for $\theta_h' \neq \theta_h$. Irreducibility and aperiodicity need to also be verified on a case-by-case basis to guarantee that π is the unique stationary distribution. However, any chain that has a positive probability of rejecting a candidate point at any iteration (and thus remaining at the same state) is guaranteed aperiodicity.

2.4.2 The Gibbs sampler

The distribution $\pi(\theta_h|\theta_{-h}^{(n-1)}, x)$, in acceptance probability (2.9), is called the *full conditional posterior distribution* for θ_h. A special case of the Metropolis-Hastings algorithm uses candidate values sampled from this distribution. Thus, at step 2 of Algorithm 2.4.1, a candidate point θ_h' is generated from the proposal distribution $q(\theta_h, \theta_h') = \pi(\theta_h'|\theta_{-h}^{(n-1)}, x)$. It is straightforward to see, from (2.9), that such a proposal (together with the corresponding one for the reverse move) results in candidate θ_h' being accepted with probability 1.

This algorithm, known as the *Gibbs sampler*, requires us to generate candidate values from the full conditional posterior distribution. In cases where the full conditional is from a standard family of distributions, Gibbs sampling is typically preferred over Metropolis-Hastings. In many applications, however, simulation from $\pi(\theta_h|\theta_{-h}^{(n-1)}, x)$ may be unfeasible or computationally expensive. In those situations, Metropolis-Hastings can always be implemented with quite an arbitrary choice of proposal distribution from which it is easy to simulate. Green and Han [43] discuss some of the advantages of Metropolis-Hastings over the Gibbs sampler in more detail.

2.4.3 Estimation and other practical issues

How can the MCMC output be used to estimate $\mathbb{E}_\pi(f(\theta)|x)$? Consider a realisation $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$ generated by an irreducible and aperiodic Markov chain that has reached its stationary distribution π. (We will suppose that we have discarded the initial M_0 iterations in which the process has not reached equilibrium yet; soon we will discuss how to estimate the value of M_0.) The *ergodic average*

$$\bar{f}_M = \frac{1}{M} \sum_{n=1}^{M} f(\theta^{(n)}) \tag{2.11}$$

is a consistent estimator of $\mathbb{E}_\pi(f(\theta)|x)$ in that it converges almost surely to the desired expectation as $M \to \infty$ (a more technical treatment can be found in Roberts [97] and Grimmett and Stirzaker [44, ch. 16]). The accuracy of the ergodic average in estimating $\mathbb{E}_\pi(f(\theta)|x)$ can be measured by the sum of the variance and the squared bias of the estimator $\bar{f}_M$, $MSE(\bar{f}_M) = Var(\bar{f}_M) + b^2(\bar{f}_M)$. The bias and the variance behave asymptotically as [43]:

$$b(\bar{f}_M) = \frac{1}{M} \sum_{n=1}^{M} \{\mathbb{E}(f(\theta^{(n)})) - \mathbb{E}_\pi(f(\theta)|x)\} \tag{2.12}$$

and

$$Var(\bar{f}_M) \sim \frac{Var_\pi(f(\theta)|x)}{M} \sum_{t=-\infty}^{\infty} \rho_t(f) \tag{2.13}$$

where $\rho_t(f)$ is the lag t autocorrelation of the stationary chain $\{f(\theta^{(1)}), \ldots, f(\theta^{(M)})\}$, which is estimated as (see for example [12]):

$$\hat{\rho}_t = \frac{\sum_{j=1}^{M-t} \left(f(\theta^{(j)}) - \bar{f}_M\right)\left(f(\theta^{(j+t)}) - \bar{f}_M\right)}{\sum_{j=1}^{M} \left(f(\theta^{(j)}) - \bar{f}_M\right)^2}.$$

Expressions (2.12) and (2.13) show that the MSE is dominated by the asymptotic variance, which is of order M^{-1} while the order of the squared bias is M^{-2}. The accuracy of estimation is thus determined by $Var(\bar{f}_M)$; the smaller the variance, the better the estimator. In practice, the complexity of π prevents explicit calculation of $Var_\pi(f(\theta)|x)$ in (2.13). However, the factor $Var_\pi(f(\theta)|x)/M$ does not depend on the Markov chain and the asymptotic variance of an estimator with respect to another is thus fully specified by the factor

$$\tau(f) = \sum_{t=-\infty}^{\infty} \rho_t(f), \tag{2.14}$$

(as long as M remains fixed): in agreement with Green and Han [43], we call $\tau(f)$ the *integrated autocorrelation time*. $\tau(f)$ provides the means for measuring how much better ($\tau(f) < 1$) or worse ($\tau(f) > 1$) a particular estimator is in comparison to independent sampling ($\tau(f) = 1$). It can be used to compare different MCMC methods: to optimise the accuracy of estimation one could choose a method with the smallest possible $\tau(f)$. However, when comparing two methods, we must not only consider the accuracy of estimation but also the computational cost. Say that method A is twice as accurate as method B but it takes six times as long to generate a certain number of samples with A than with B. Then method A is, in effect, less cost-efficient than B and method B should then be preferred.

In our implementation, we use Geyer's [36] *initial positive sequence estimator* to estimate $\tau(f)$, given by:

$$\hat{\tau}(f) = -1 + 2\sum_{i=0}^{K} \hat{\Gamma}_i \tag{2.15}$$

where $\hat{\Gamma}_i = \hat{\rho}_{2i} + \hat{\rho}_{2i+1}$ is the sum of adjacent pairs of sample autocorrelations and $\hat{\rho}_t$ is the autocorrelation at lag t. Here K is chosen to be the largest integer such that $\hat{\Gamma}_i > 0$ (for $i = 0, 1, \ldots, K$).

The integrated autocorrelation time encodes the information about the correlation structure of the chain; the greater the correlation between the samples, the larger the $\tau(f)$. It is in this sense that $\tau(f)$ is closely related to the term *mixing* as defined by Geyer [36]. He refers to 'mixing' as the dependence between samples at lag t, as measured by their correlation. A chain that mixes quickly moves more agilely around the support of π (and therefore has a smaller $\tau(f)$) than a chain that mixes slowly. In this sense a rapidly mixing chain produces more reliable estimates of $\mathbb{E}_\pi(f(\theta)|x)$ than a slowly mixing one, for fixed sample size M, and the former is usually preferred over the latter (as long as the computational cost of the rapidly-mixing chain is not prohibitive). In this study, we use the term 'mixing' in agreement with Geyer's [36] interpretation.

So far the discussion has focused on a scenario in which the Markov chain has already converged to the stationary distribution. Global balance ensures that once a sample from the stationary distribution has been obtained, all subsequent samples will be from that distribution. The issue is thus to construct a chain that converges reasonably fast to stationarity and that, once having converged, has good estimation performance (small $\tau(f)$). Green and Han [43] investigated strategies for achieving both, and suggested a *combined transition mechanism* in which an initial chain is used for the first M_0 iterations then switching to a different chain for another M steps. The first chain is chosen to give rapid convergence to stationarity and the second for small $\tau(f)$. Implementation of such a mechanism would require monitoring convergence to stationarity as the simulation proceeds; when diagnostics suggest that the process has reached equilibrium, the switch would take place. In this thesis we have not attempted a combined transition strategy of this form.

It is common practice in MCMC simulation to discard the initial M_0 iterations. These initial iterations are referred to as the *burn-in* period and they usually correspond to a stage where the chain has not reached equilibrium yet. There exist formal tools for estimating the length of the burn-in (for an overview see [38]). Our approach involves the visual inspection of the plots of (functions of) the MCMC output against iteration number. M_0 is set to a value at which the trace of samples exhibits stability. In most of our analyses, discarding the first quarter of the samples has been enough.

The length of the run is chosen depending on the computational cost of the simulation and on the information about convergence to stationarity and estimation performance provided by initial exploratory runs. In all our analyses we required a number of exploratory runs before deciding on a suitable value for the run length.

Finally, some considerations about the starting value for the chain. In the type of Markov chains we are concerned with, the choice of starting value will not affect the ultimate convergence of the chain to the stationary distribution. A rapidly mixing chain will quickly find its way from extreme starting values while a slowly mixing one may need a longer burn-in. In most cases, a starting point sampled at random from the prior distribution is acceptable. In some complex situations, however, the tree space is so vast that even a rapidly mixing chain would take very long to find regions of high posterior support. In cases like this, we have taken the approach of initialising the tree at its maximum likelihood estimate.

2.4.4 Improving MCMC by tempering

There exist complex problems in which the target distribution contains several modes. The MCMC sampler then has to be able to move between the modes in order to produce an adequate sample from the entire target distribution. It may occur that the modes are separated by regions of extremely low probability density so that the chain may find it difficult to 'escape' from the neighbourhood of a certain mode. One strategy for improving the performance of the sampler in this situation is to sample from a distribution $\pi_1(\theta|x)$ that is a 'flattened' version of the original target $\pi_0(\theta|x)$. The problem then becomes to decide how much to flatten $\pi_0(\theta|x)$. If π_1 and π_0 are not very different, the improvement is not much as π_1 may not be that much easier to sample from than π_0. On the other hand, if the two distributions are different enough that the problem of modality is overcome, they may also be different enough that π_1 is no longer a reliable representation of the target π_0. The basis of tempering methods is the introduction of a series of *tempered distributions* bridging the gap between π_0 and π_1 [3]. Gelman, Carlin, Stern and Rubin [32, ch. 13], and Gilks and Roberts [39] discuss in more detail some criteria for choosing these tempered distributions. For our purposes, all that matters is that a series of successively easier-to-sample distributions are introduced into the MCMC sampler to help improve the movement of the chain around the state space. One strategy for incorporating these tempered distributions into the sampler is to run $K+1$ parallel chains, each with its own target distribution $\pi_k(\theta|x)$, $k = 0, 1, \ldots, K$, where $\pi_1, \ldots, \pi_K$ are tempered versions of π_0. At each iteration, each chain is updated separately and an additional Metropolis-Hastings step is used to attempt to swap the states of the main chain with one of the tempered chains. At the end of the run, the output from the tempered chains is discarded and only the main chain is utilised to draw inferences. This method is called *Metropolis-coupled MCMC* [35]. There exist other tempering methods, such as *simulated tempering* [72, 37] and *tempered*

transitions [84]. This section does not intend to provide an exhaustive account of tempered MCMC but rather to highlight its high computational cost. In Metropolis-coupled MCMC for example, the computational effort is K times more than that of an ordinary MCMC run. In the following, whenever we refer to tempered MCMC, we will mean Metropolis-coupled MCMC.

2.5 Continuous-time Markov processes

Let $\{\theta(t) : t \geq 0\}$ be a collection of random variables taking values in some discrete state space Θ and indexed by the set $[0, \infty)$. The process $\{\theta(t) : t \geq 0\}$ is called a *continuous-time Markov process* if it satisfies the condition

$$Pr(\theta(t_{n+1}) = j|\theta(t_0) = i_0, \ldots, \theta(t_n) = i) = Pr(\theta(t_{n+1}) = j|\theta(t_n) = i) \tag{2.16}$$

for all states $i_0, \ldots, i, j \in \Theta$ and any sequence $t_0 < \ldots < t_n < t_{n+1}$ of times. As before, we concentrate exclusively on processes where the laws of evolution are time invariant, that is $Pr(\theta(t) = j|\theta(s) = i) = Pr(\theta(t-s) = j|\theta(0) = i)$ for $i, j \in \Theta$ and $s \leq t$. If this holds, the process is called a *homogeneous Markov process*.

The *transition probability* $p_{ij}(t)$ of a homogeneous Markov process is defined as

$$p_{ij}(t) = Pr(\theta(t) = j|\theta(0) = i), \quad \text{for } i, j \in \Theta. \tag{2.17}$$

If we write $P(t)$ for the $|\Theta| \times |\Theta|$ matrix with entries $p_{ij}(t)$, then $P(t)$ is called the *transition probability matrix* of the Markov process. The rules of evolution of the process are contained in $P(t)$; entry $p_{ij}(t)$ indicates the probability that the process, in state i, jumps to state j over a period of time t. The transition probabilities satisfy $p_{ij}(t) > 0$ and $\sum_{j \in \Theta} p_{ij}(t) = 1$, for all $i \in \Theta$ and $t \geq 0$. In addition, they satisfy the Chapman-Kolmogorov equation:

$$\sum_{k \in \Theta} p_{ik}(s) p_{kj}(t) = p_{ij}(s+t) \quad \text{for all } s, t \geq 0 \tag{2.18}$$

and the initial condition $p_{ij}(0) = 1$, if $i = j$ and $p_{ij}(0) = 0$, if $i \neq j$.

The transition probability matrix $P(t)$ is computed from a *generator matrix* Q by exponentiation of the form

$$P(t) = e^{Qt}. \tag{2.19}$$

Exponentiation of generator matrices in phylogenetic analyses is not always straightforward and Section 3.4.1 discusses how to compute phylogenetic transition probability matrices.

The entries in the generator, or rate, matrix Q satisfy:

(i) $0 \leq -q_{ii} < \infty$ for all $i \in \Theta$;
(ii) $q_{ij} \geq 0$ for all $i \neq j,\ i, j \in \Theta$;
(iii) $\sum_{j \in \Theta} q_{ij} = 0$ for all $i \in \Theta$.

Entry q_{ij} indicates the rate at which the process enters state j when in state i ($i \neq j$) and entry q_{ii} gives the total rate at which the process remains at state i. There are several Q matrices of fundamental importance in phylogenetic applications, some of which are presented in Chapter 3.

The long-term behaviour of the process is closely related to the existence of stationary distributions. The vector $\nu = \{\nu_j :\ j \in \Theta\}$ is called a *stationary distribution* of the Markov process if $\nu_j \geq 0$, $\sum_j \nu_j = 1$ and global balance is satisfied with respect to the transition $p_{ij}(t)$, for all $t \geq 0$ (Section 2.3.3).

2.6 Finite mixture models

The basic idea behind mixture modelling is to assume that an observation arises from a superposition of k generating distributions. Because these component distributions are usually taken to be simple distributions, mixture models provide a convenient way of modelling quite complex non-standard distributions.

2.6.1 Basic formulation

In mixture modelling, we assume that observations $x_1, \ldots, x_N$ arise from a distribution of the form

$$x_n \sim \sum_{j=1}^{k} \omega_j p_j(x_n|\theta_j), \quad \text{independently for } n = 1, \ldots, N$$

where $\omega_1, \ldots, \omega_k$ are nonnegative quantities that sum to one, called the weights or *mixture proportions*, and $\theta_1, \ldots, \theta_k$ are k distinct (possibly vector) parameters that index the *component distributions* $p_1, \ldots, p_k$, respectively. In this thesis, we assume that the component distributions belong to the same parametric family so that

$$x_n \sim \sum_{j=1}^{k} \omega_j p(x_n|\theta_j), \quad \text{independently for } n = 1, \ldots, N$$

where $p(\cdot|\theta)$ denotes a generic member of the parametric family $\{p(x_n|\theta) : \theta \in \Theta\}$. In the context of phylogenetic analyses, each of the k components of a mixture may have a direct biological interpretation. For instance, one component may conform to data arising from one gene while another component may describe the evolution at a different gene. We

may thus think of each observation as arising from one of the k components. The identity of the component from which each observation is drawn is unknown and we thus regard the component identity for that observation as a missing variable. This interpretation of a mixture is appropriate in terms of classification of the observations and is the one adopted in this thesis. It requires us to reformulate the mixture in a 'missing-data' way.

2.6.2 Missing-data reformulation

In a missing-data reformulation of a finite mixture model, observation x_n is 'augmented' by a variable z_n. Quantity z_n, called the *allocation variable* of x_n, is an unobserved integer which takes values in the set $\{1, \ldots, k\}$ and identifies the underlying generating component of observation x_n. Conditional on z_n, observation x_n is independently drawn from the distribution corresponding to the z_nth mixture component. That is,

$$x_n | z_n \sim p(x_n | \theta_{z_n}), \quad \text{independently for } n = 1, \ldots, N. \tag{2.20}$$

Expression (2.20) says that, once the allocation for x_n is known, this observation is no longer considered as generated by a mixture of k distributions but as generated by the z_nth component distribution. This approach has long been used in the statistical literature when one of the objectives of the analysis is to classify a set of observations (e.g. [122, 4]) and is extensively employed in this thesis.

2.7 Phylogenetic data

Our presentation concentrates on DNA data, also known as nucleotide sequence data. DNA is the agent responsible for carrying the hereditary information in almost all living organisms. It is a large macromolecule consisting of two complementary strands twined around each other [120]. Each strand is a sequential arrangement of four types of basic molecules called *nucleotides*. Each nucleotide contains a phosphate group, a sugar (deoxyribose) and one of four *bases* – adenine (A), cytosine (C), guanine (G) and thymine (T) [87]. DNA material is said to be *sequenced* when the succession of bases that form one of the two complementary strands is determined. The alphabet of a *DNA sequence* is thus a set of four letters that correspond with the four constituent bases. Henceforth, we refer to the characters of a DNA sequence as nucleotides.

Our analysis starts by observing the *aligned* DNA sequences of a set of organisms. Figure 2-2 shows a *DNA alignment* of five strains of the *Borrelia burgdorferi* bacterium, one of the bacterial species responsible for Lyme disease. The sequences in this alignment correspond to (a portion of) the gene that codes for the outer surface protein C, called the *ospC* gene. This alignment contains 54 position in which all strains share the same character and 6 in which characters differ (highlighted as bold letters). Positions of the former type are called *monomorphic* while positions that exhibit variation in nucleotides are

```
B31      TCTGCTGATGAGTCTGTTAAAGGGCCTAATCTTACAGAAATAAGTAAAAAAATTACGGAT
IPT2     TCTGCTGATGAGTCTGTTAAAGGGCCTAATCTTACAGAAATAAGTAAAAAAATTACGGAT
IPT19    TCTGCTGATGAGTCTGTTAAAGGGCCTAATCTTGCAGAAATAAGTAAAAAAATTACAGAA
Z41293   TCTGCTAATGAGTCTGTTAAAGGGCCTAATCTTACAGAAATAAGTAAAAAAATTACAGAA
Z41493   TCTGCTGATGAGTCTGTTAAAGGACCCAATCTTACAGAAATAAGTAAAAAAATTACAGAT
```

Figure 2-2: Alignment of 60 nucleotides from five strains of *Borrelia burgdorferi*. The DNA sequences correspond to a portion of the gene that encodes the outer surface protein C. The alignment contains 6 *polymorphic* positions, which have been highlighted.

called *polymorphic*. Either of the monomorphic or polymorphic type, each column in the alignment is called a *site*.

A requirement for meaningful phylogenetic inference is that the DNA sequences are *homologous*, which in general means *inferred* common ancestry [83]. Suppose that two different species share a gene that was acquired by direct descent from a common ancestor and that a certain functionality of the gene is hypothesised to have already existed in the ancestor. Then we say that the two genes are homologous for that functionality. However, if that functionality was acquired independently by the two species then that functionality is not homologous. An example is found in the lysozyme gene of cows and leaf-eating langur monkeys. This gene, in its 'conventional form', was passed on to the two animals by a common ancestor [87]. Both animals, however, independently evolved a digestive function of the lysozyme gene to become ruminants. Therefore, langur and cow lysozymes are homologous as conventional genes but non-homologous as digestive enzymes (because the latter functionality was absent in the ancestral lysozyme). Sequence homology is a way of ensuring that the functionality of the analysed sequences can be traced back to a similar biological function that existed in the common ancestor. This way, it makes sense hypothesising evolution by common ancestry and thus reconstructing the phylogenetic history that relates the organisms of interest.

Even if two DNA sequences are hypothesised to be homologous, yet another level of homology is necessary for meaningful nucleotide-based phylogenetic inference; homology of individual sites. That is, the characters observed at a given position in the alignment should all trace their ancestry back to a single position that occurred in the common ancestor of all analysed organisms. Inferring such *positional homology* is the purpose of *sequence alignment* methods, an area outside the scope of this thesis. In our analyses, we assume that the sequences have been suitably aligned to allow for valid phylogenetic conclusions. Moreover, in the DNA alignments that we examine all sites containing *gaps* (i.e. characters '-' as opposed to 'A', 'C', 'G' or 'T') or other ambiguous characters have been removed. A thorough account of sequence alignment can be found in the textbook by Haubold and Wiehe [47]. Moritz and Hillis [83] discuss in more detail the concept of homology as so Page and Holmes [87] do.

2.8 Terminology

2.8.1 The evolutionary tree and its branch lengths

An evolutionary, or phylogenetic, tree is a structure that can be defined formally by borrowing some concepts from graph theory. A graph is a collection of *vertices* that can have connections between them, called *edges*. The edges may be weighted by positive real numbers, called the *edge weights*. An *undirected graph* has edges that can be traversed in either direction, from vertix v to w or from w to v, where v and w are two members of the set of vertices V that are connected by an edge. A *path* is a sequence of vertices that describes a route along edges through the graph. An undirected graph is said to be *connected* when there is a path, using any number of edges, from any one vertex to all other vertices. A graph containing no paths that begin and end at the same vertex is said to be an *acyclic graph*. A *tree* T is a connected acyclic graph. The *degree* of a vertex is the number of edges which touch it and a vertex of T with degree one is called a *leaf*. The set of all the leaves of the tree is denoted by L ($L \subset V$). A vertex which is not a leaf is called an *interior vertex*. In this thesis we only consider trees whose interior vertices have all degree three, called *strictly bifurcating trees*.

In phylogenetic parlance, edges are usually referred to as *branches*, edges weights are called *branch lengths* and vertices are called *nodes*. The nodes, branches and branch lengths symbolise real objects or states. For instance, the interior nodes represent ancestral organisms while the leaf nodes stand for extant entities. The branches represent the evolutionary processes that link the organisms and the branch lengths represent the amount of evolutionary divergence between the organisms. (We will return to the issue of the interpretation of branch lengths in Section 3.4.2.) The length of a branch is always measured as the horizontal distance between two nodes (see Figure 2-3(a)).

Let A be a non-empty finite set with size equal to the cardinality of L, and let the map $\varphi : A \rightarrow L$ be a bijection. Then $\phi = (T, \varphi)$ is called a *phylogenetic tree on* A with *labelling function* φ and *label set* A [105]. The label set contains the names that identify the organisms under analysis. To illustrate this, consider the DNA alignment of the *B. burgdorferi* bacterium in Figure 2-2. The label set in this example is $A = (B31, IPT2, IPT19, Z41293, Z41493)$ and Figure 2-3(a) shows a phylogenetic tree on this label set. (In this thesis we interchangeably represent a phylogenetic tree as either of the two trees in Figure 2-3(a).) Sometimes we are not interested in the specific branching structure of a subtree of a phylogenetic tree, and we simply collect the entire subtree in a triangular node as in Figure 2-3(b). In this case, the triangular node indicates that there is a subtree with two leaf-nodes (labelled as $IPT2$ and $IPT19$) attached to that point of the tree.

There is a fundamental distinction between the *state* and the *label* of a node. In the case of a leaf node, its state is determined by an observable DNA character but its label is

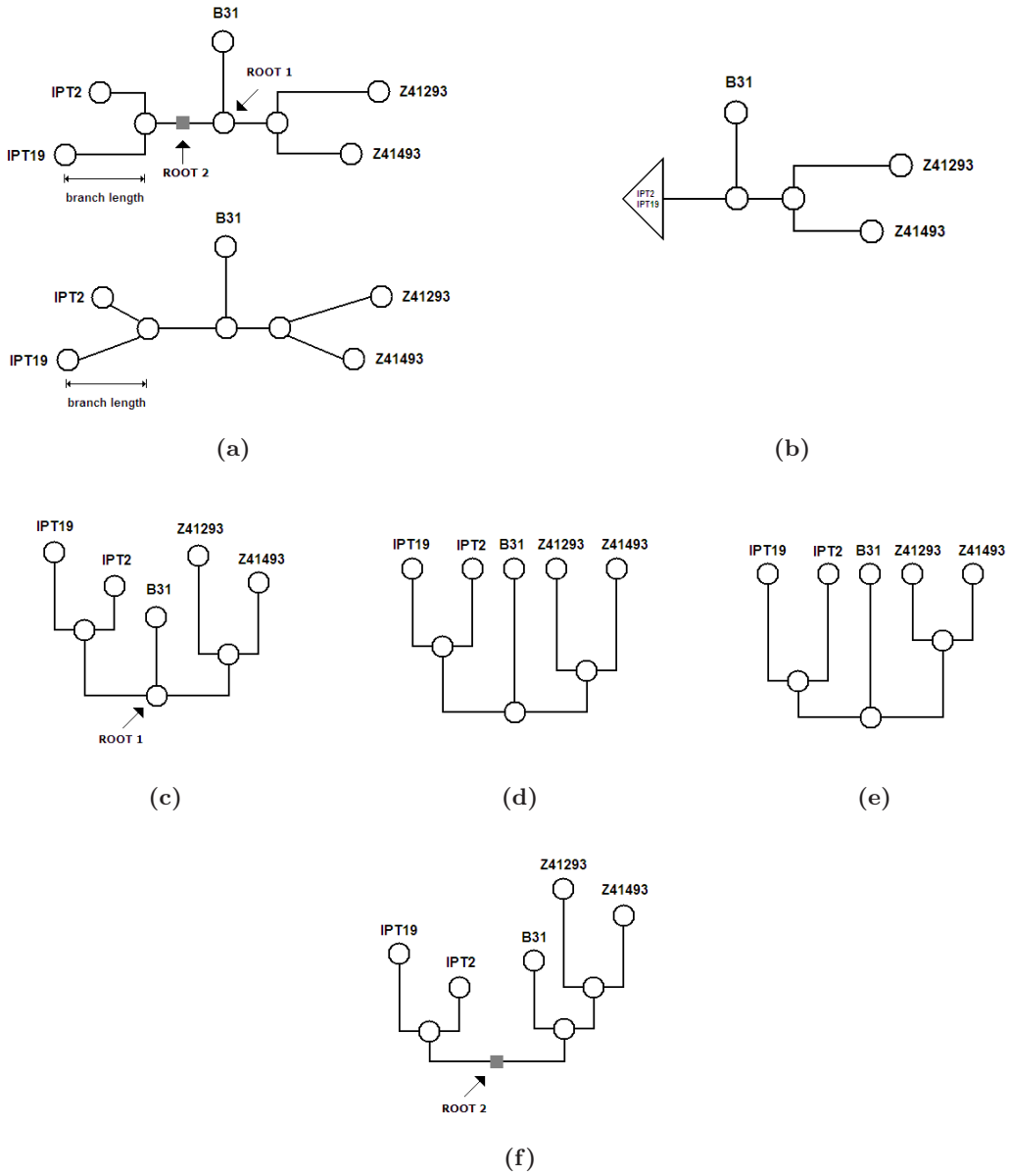


Figure 2-3: (a) A *phylogenetic tree* on the label set $(B31, IPT2, IPT19, Z41293, Z41493)$ with the length of a branch indicated, and an equivalent representation of the same tree underneath. (b) A representation of a phylogenetic tree in which one entire subtree is collected in a triangular node, meaning that the specific branching structure of the subtree is not relevant. (c) The phylogenetic tree from (a) arbitrarily rooted at one of its interior nodes. (d) A *clock-like tree* in which all the leaf nodes are placed at the same distance from the root. If the order of the interior nodes with respect to the root is recorded, this tree is called a *labelled-history*. (e) A different labelled-history to that in (d) since the order of occurrence of the interior nodes relative to the root is distinct. (f) A *strictly bifurcating rooted tree* with its 'artificial' root indicated as a grey square.

assigned from the label set A. (None of Figures 2-3(a)-(f) show the states of the nodes, but only the labels.) Throughout, we will always write the state of a node as a single character enclosed by the node itself, while the label will appear next to the node. In contrast with the leaf nodes, which are uniquely labelled by the set A, the interior nodes are arbitrarily labelled. Consequently, more than one distinct labelling of interior nodes may map to the same phylogenetic tree on a label set A.

It is possible to root a phylogenetic tree by simply distinguishing any one node, either interior or exterior. We call this 'distinguished node' the *root* of the tree. Thus Figure 2-3(c) shows the same phylogenetic tree as in (a) but this time arbitrarily rooted at one of its interior nodes. Some of the methods discussed in this thesis focus attention on trees with all their leaves lying at the same distance from the root, called *clock-like trees*. Figure 2-3(d) shows a clock-like tree on the label set A. Moreover, there exist applications that use clock-like trees in which the order of the interior nodes with respect to the root is recorded, called *labelled histories*. Thus Figures 2-3(d) and (e) show the same clock-like tree but different labelled histories as the order of the interior nodes relative to the root is different. That is, the labelled history in Figure 2-3(d) has the interior node that joins leafs $Z41293$ and $Z41493$ closer to the root than the interior node that joins leafs $IPT2$ and $IPT19$. In contrast, in Figure 2-3(e) this order has been inverted.

Some phylogenetic methods restrict attention to strictly bifurcating *rooted* trees, which means that the root node has degree two, the interior nodes degree three and the leaf nodes degree one. The tree in Figure 2-3(c) is not a strictly bifurcating rooted tree because the root node has degree three instead of two. It is possible to root an unrooted tree in a 'strictly bifurcating way' by inserting an 'artificial' node at any branch, like the node indicated as '$\mathrm{Root}\ 2$' in Figure 2-3(a). The resulting strictly bifurcating rooted tree is shown in Figure 2-3(f). In this thesis, we will make interchangeable use of 'arbitrarily rooted trees at any node' as in Figure 2-3(c), or 'strictly bifurcating rooted trees' as in Figure 2-3(f). We will later demonstrate that our model ultimately returns an unrooted rather than a rooted tree. Therefore, all that matters to our method is that the unrooted tree is strictly bifurcating. For instance, all the trees shown in Figures 2-3(c)-(f) map to the strictly bifurcating *unrooted* tree displayed in Figure 2-3(a) and our method does not distinguish between them. In the following, the reader should assume that a tree is unrooted unless otherwise specified.

A last consideration is that phylogenetic trees are called in a number of different ways in the literature: evolutionary trees, phylogenies, tree topologies or, simply, trees. In this thesis we make use of these names interchangeably. Also, we generically refer to the organisms analysed in a phylogenetic study as *taxa*.

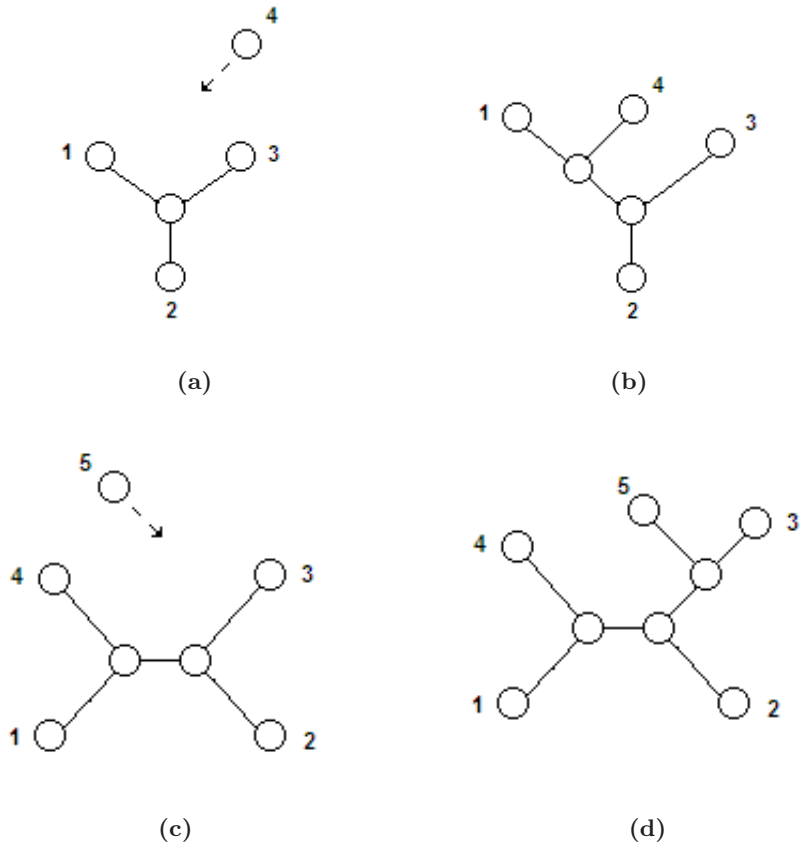


Figure 2-4: (a) A three-leaf tree has three branches to which a forth node could be added to produce a tree of four leaves, like in (b). (c) A four-leaf tree has five branches into which a fifth node could be inserted to generate a five-leaf tree, like in (d).

2.8.2 How large is the tree space?

Let $\Phi(S)$ be the set of all strictly bifurcating phylogenetic trees on the label set $(1, 2, \ldots, S)$. It is possible to calculate the cardinality of $\Phi(S)$ by the following argument [10]: starting with $S = 3$, there is only one way in which three labels can be assigned to the leaves of a strictly bifurcating, unrooted phylogenetic tree, hence $|\Phi(3)| = 1$. This small tree has three branches to which a fourth node could be added (see Figures 2-4(a) and (b)), which results in three possible ways of generating a tree of four leaves, or $|\Phi(4)| = 3$. It is thus equivalent to write $|\Phi(4)| = |\Phi(3)| \times$ (number of branches of a three-leaf tree). A four-leaf tree has five branches into which a fifth node could be inserted (see Figures 2-4(c) and (d)), thus $|\Phi(5)| = |\Phi(4)| \times$ (number of branches of a four-leaf tree) $= 15$. Each branch of a tree is the location of a possible leaf-node addition and, in order to compute the number of trees on a label set $(1, 2, \ldots, S)$, it is necessary to first find an expression for the number of branches of an S-leaf tree. Such an expression can be obtained from the following two propositions.

Proposition 2.8.1. *Let ϕ be a strictly bifurcating phylogenetic tree on the label set $(1, 2, \ldots, S)$. Then, for $S \geq 2$, tree ϕ has $|V|_S := 2S - 2$ nodes.*

Proof. The proof is by induction on S. Since the number of nodes of a strictly bifurcating phylogenetic tree on $(1, 2)$ is $|V|_2 = 2$, the result holds for $S = 2$. Assume the result is true for $S = k$. By adding an extra leaf to the tree, the size of the set of nodes V is incremented by two; the new leaf plus a new node that appears at the point of incidence of the branch that joins the new leaf with ϕ. This is shown in Figure 2-5. The number of nodes for $S = k + 1$ may then be written as $|V|_{k+1} = |V|_k + 2$. Therefore, by the induction assumption,

$$\begin{aligned} |V|_{k+1} &= 2k + 2 - 2 \\ &= 2(k + 1) - 2 \end{aligned}$$

as required for $S = k + 1$.

□

Proposition 2.8.2. *The number of branches of a strictly bifurcating phylogenetic tree on the label set* $(1, 2, \ldots, S)$ *is* $|E|_S := |V|_S - 1$*, for* $S \geq 2$.

Proof. Once again, this proof is by induction on S. Evidently, this proposition holds for $S = 2$ since there is only one branch in a two-leaf tree and $|V|_2 - 1 = 1$. Suppose that the implication holds for $S = k$. The number of branches for $S = k + 1$ may be written as $|E|_{k+1} = |E|_k + 2$, since two new branches are added to the existing tree by including an extra leaf, namely, the branch that connects the new leaf with the tree plus an extra branch that results from splitting an existing branch at the place where the new branch is connected, as shown in Figure 2-5. Therefore, by the induction assumption and by knowing that $|V|_k = |V|_{k+1} - 2$ (see the proof to Proposition 2.8.1), it is possible to write

$$\begin{aligned} |E|_{k+1} &= |E|_k + 2 \\ &= |V|_k + 1 \\ &= |V|_{k+1} - 1 \end{aligned}$$

which is the desired result for $S = k + 1$.

□

According to Propositions 2.8.1 and 2.8.2, the number of branches on an S-leaf phylogenetic tree may be written as $|E|_S = |V|_S - 1 = 2S - 3$. It is thus possible to find an expression for the size of the space of all phylogenetic trees on the label set $(1, 2, \ldots, S)$ as follows:

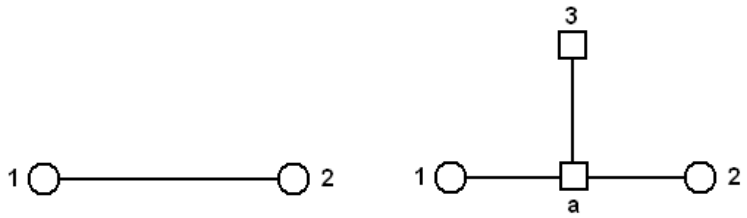


Figure 2-5: Beginning with a phylogenetic tree on the label set $(1, 2)$, the addition of a new leaf (in this case the node with label 3) will result in the set of nodes changing from $\{1, 2\}$ to $\{1, 2, 3, a\}$. Similarly, the original set of branches, which only contained the branch that connects nodes 1 and 2, becomes a set containing the branches joining nodes 1 and a, 2 and a, and 3 and a.

$$\begin{aligned} |\Phi(S)| &= |\Phi(S-1)| \times (\text{number of branches of a } (S-1)\text{-leaf tree}) \\ &= |\Phi(S-1)| \times |E|_{S-1} \\ &= 1 \times 3 \times 5 \times \ldots \times (2S-5) \\ &= (2S-5)!/(2^{S-3}\,(S-3)!). \end{aligned} \tag{2.21}$$

Expression (2.21) reveals the vastness of the tree space. In a phylogenetic analysis with as few as 10 taxa, there are over 2×10^6 phylogenetic trees. The immediate implication of these large numbers is that dealing with trees is computationally expensive and the use of efficient algorithms is essential.

2.9 Software implementation

MCMC methods devised as part of this research have been implemented by the author of this thesis in the C program Arbol. The program takes as input a DNA alignment and the values of some execution parameters (such as the length of the run and the length of burn-in) and it returns as output the sample path of the simulated Markov chain. Additional programs for summarising the MCMC output and estimating integrated autocorrelation times were also developed.

Chapter 3

The homogeneous phylogenetic model

Statistical inference of phylogenies almost didn't happen.
– Joseph Felsenstein

3.1 Introduction

This chapter discusses the details of the conventional probabilistic model for phylogenetic inference, which we have called the *homogeneous phylogenetic model*. This chapter starts by presenting a brief account of the development of statistical methods for phylogenetics and continues by motivating the process that we are interested in modelling; the evolution of an ancestral DNA sequence into extant sequences on a very large time-scale. Firstly, we use rooted trees to represent evolution as a process that begins at the root of the tree and proceeds towards the leaves, following bifurcating paths that indicate how ancient organisms give rise to two descendant entities. The root in this tree represents the common ancestor to all organisms and the leaves symbolise extant taxa. Viewing evolution as a 'rooted process' is intuitive, easy to interpret and it simplifies the discussion of the model at first. However, the homogeneous model can only return an unrooted rather than a rooted tree (at least directly), and in this chapter we examine the impossibility of inferring rooted trees.

The homogeneous phylogenetic model is formulated in terms of three component-parts: a tree parameter, the branch lengths of this tree and the parameters of an infinitesimal Q-matrix that generates a continuous-time Markov process. Both tree and branch lengths have been discussed before. In this chapter, we introduce a Markov process as the means to model the evolution of DNA characters through time; from ancient DNA sequences whose characters are substituted to give rise to descendant sequences. There are very many different Q-matrices that generate a process like this. In this chapter, we present some of the most well-known of these matrices.

The computation of the likelihood function for phylogenetic parameters is not straightforward due to the rather non-standard form of the homogeneous model. In this chapter, we discuss the details of the calculation of likelihood phylogenetic functions and review a popular methodology for the efficient calculation of this function, called the 'pruning algorithm' [23]. The remainder of this chapter presents details on the choice of prior distributions for model parameters and finally, an overall discussion.

3.2 Background

The statistical inference of phylogenetic trees based on likelihood methods was introduced by Edwards and Cavalli-Sforza in 1964 [19]. Given the unavailability of DNA sequences at the time, they used gene frequency data to reconstruct the evolutionary tree that related fifteen human populations. They thought of evolution as a branching random walk, with a constant probability of branching and a constant rate of walking. They attempted a maximum-likelihood approach but reported inadequacy of their computer program and failed to obtain valid parameter estimates.

Some years later, in 1971, Jerzy Neyman published a paper with the purpose of bringing to the attention of the community of statisticians 'a source of novel statistical problems', as he referred to the study of evolution from molecular data [85]. He based his analyses on DNA data, and used a formulation that assumed independence of sites and that modelled each of them with the same mechanism of evolution – assumptions that remain to this day.

Joseph Felsenstein, in his 1973 paper on the maximum-likelihood estimation of evolutionary trees from DNA data, used a Markov process to model the evolution of characters along the branches of a tree [22]. He got rid of Edward and Cavalli-Sforza's branching process and considered, instead, the tree graph as a model parameter to be estimated. This greatly simplified the mathematics and fostered the growth of statistical phylogenetics based on likelihood methods. In 1981, Felsenstein showed how to make likelihood computations for DNA data practical for an arbitrary number of sequences [23], a method called the 'pruning algorithm '. Before this, the calculation of the likelihood function was cumbersome and only convenient for moderately-sized problems. Felsenstein's legacy to the field is of great value and most of the methods discussed in this thesis can be considered direct descendants of his likelihood approach.

Rannala and Yang, in 1996, attempted one of the first ever Bayesian approaches to the phylogenetic problem [93]. In their work, they only performed Bayesian inference on the phylogenetic tree and estimated all other parameters by frequentist techniques, prior to calculating the posterior distribution for trees. To evaluate the posterior, Rannala and Yang used numerical integration, which turned out to be impractical for more than five DNA

sequences. One year later they expanded their work by introducing MCMC methods [131], but more details on this belong to the next chapter.

The adoption of rigorous statistical methods for the inference of phylogenies has not been straightforward. Felsenstein [26] tells the story of how statistical phylogeneticists had to prove wrong those advocates of more philosophical approaches; effort that took them decades. Nowadays, the use of likelihood-based methods for the reconstruction of evolutionary histories is common and interest in the field continues to grow.

3.3 Modelling the evolution of organisms

We are interested in modelling the evolution of an ancestral (unobserved) DNA sequence into a number of descendant (observed) sequences, operating on a very large time-scale. The evolutionary process starts at the individual level. A change in DNA material takes place in a specific individual. If such a change is advantageous (and assuming evolution by natural selection), the organism will leave more offspring who in turn leave even more offspring, which ultimately causes the advantageous information to spread throughout the population. If the change is disadvantageous, the organism will vanish from the population. In this problem, a single sampled DNA sequence is considered as representative of the genetic make-up of an entire population (e.g. a representative of a species, or of a bacterial strain) as, presumably, all individuals within that population carry (virtually) identical genetic information. Then, for example, we draw phylogenetic conclusions about the entire human population based on one sampled human DNA sequence.

Consider the evolution of three species $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{T}$ that stem from a common hypothetical ancestor $\mathcal{Z}$. Assume that species $\mathcal{Z}$ evolves for some time before splitting into two entities, say $(\mathcal{S})$ and $(\mathcal{R}\,\mathcal{T})$. Next, species $(\mathcal{S})$ and $(\mathcal{R}\,\mathcal{T})$ continue to evolve *independently* until the ancestral lineage $(\mathcal{R}\,\mathcal{T})$ splits into $(\mathcal{R})$ and $(\mathcal{T})$, yielding the tree in Figure 3-1(a). This tree is one of three possible trees that could be formed by the bifurcation of $\mathcal{Z}$; evolution could have similarly led to a split $(\mathcal{R})$ and $(\mathcal{S}\,\mathcal{T})$, or $(\mathcal{T})$ and $(\mathcal{R}\,\mathcal{S})$. In either case, the tree would have the same branching structure, the only difference would be in the assignment of labels to the leaves.

If the number of species is larger than three, the number of possible trees increases. The first split of a common ancestor of four species, for example, may be either into just one of them $(\mathcal{R})$ plus the other three $(\mathcal{S}\,\mathcal{T}\,\mathcal{U})$, or into two pairs of ancestors $(\mathcal{R}\,\mathcal{T})$ and $(\mathcal{S}\,\mathcal{U})$, as illustrated in Figures 3-1(b) and (c). But these figures show only one of several different possibilities; ancestor $\mathcal{Z}$, in Figure 3-1(b), could have instead split into $(\mathcal{T})$ and $(\mathcal{R}\,\mathcal{S}\,\mathcal{U})$. Depending on the splitting pattern followed by the species, the leaves of the tree will be labelled in one way or another. It was Felsenstein who, in 1981, first suggested taking the *leaf-labelled* tree to be an unknown parameter and not to attempt using a probabilistic

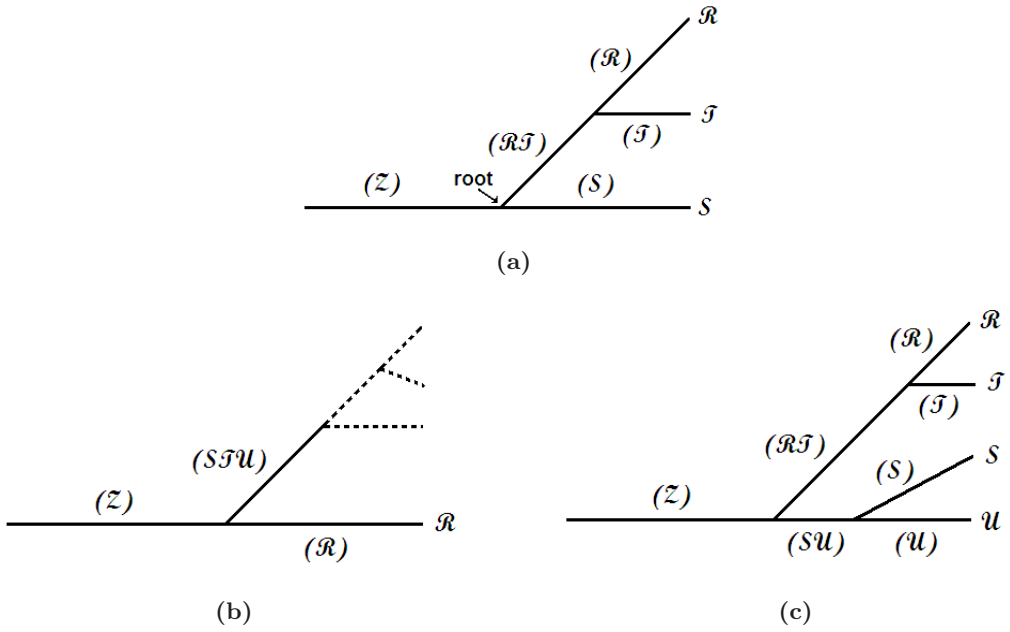


Figure 3-1: (a) Phylogenetic tree for three species, $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{T}$, descending from a common hypothetical ancestor $\mathcal{Z}$. (b) One of the possible trees for four species in which an hypothetical common ancestor splits into entities $(\mathcal{R})$ and $(\mathcal{STU})$. An alternative tree, one resulting from a split into two pairs of ancestors, is displayed in Figure (c), where the split was into $(\mathcal{RT})$ and $(\mathcal{SU})$. In either figure, a different split of the species would lead to a different labelling of the leaves.

model of the branching process that generates this tree [23]. This is the approach that we will follow in this thesis. By estimating the leaf-labelled tree we are in effect reconstructing the evolutionary relationships that hold between a set of organisms (as supported by the observed data).

The observations for our analysis are the individual sites in the DNA alignment. Suppose that the DNA alignment for species $\mathcal{R}$, $\mathcal{S}$ and $\mathcal{T}$ (from Figure 3-1(a)) is

$$\boldsymbol{x} = \begin{pmatrix} A & C & A & A & G & G & A & \cdots & T \\ G & C & A & C & G & G & C & \cdots & T \\ A & C & C & A & G & G & A & \cdots & G \end{pmatrix}$$

In this alignment, the first row corresponds to the DNA sequence of species $\mathcal{R}$, the second row to $\mathcal{S}$ and the third row to species $\mathcal{T}$. The observations are then given by $x_1 = (A, G, A)^T$, $x_2 = (C, C, C)^T$, ..., $x_N = (T, T, G)^T$. Each observation is a vector with as many elements as the number of organisms under study S (in this case, $S = 3$). The homogeneous phylogenetic model postulates that all sites in this alignment evolve independently *under the same process of evolution*. The assumption of independence allows us to concentrate on modelling the evolutionary process at a single site.

Consider the first observation in the alignment above, $x_1 = (A, G, A)^T$. Figure 3-2(a)

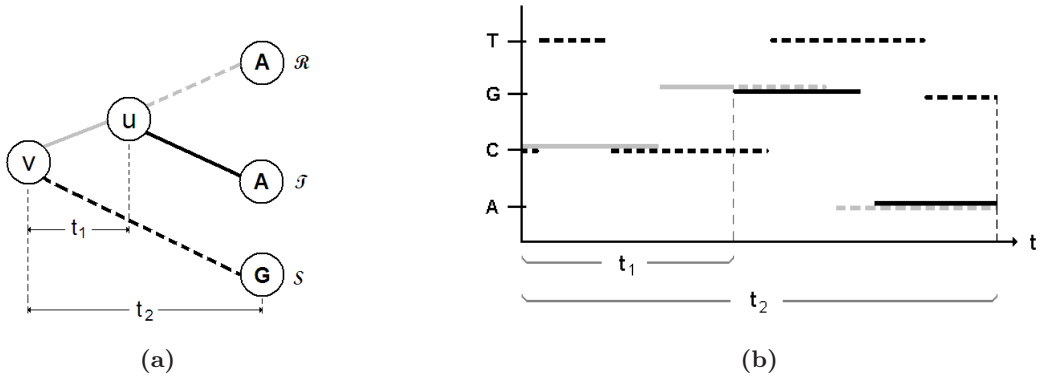


Figure 3-2: (a) A tree that relates the characters at site $x_1 = (A, G, A)^T$. (b) A sample path of the Markov process of nucleotide substitution dictated by the tree topology and branch lengths shown in (a).

shows a tree relating these characters. (Notice how the observed characters are placed at the leaves, and the characters at the interior nodes are unknown and denoted by u and v.) This tree tells us that species $\mathcal{R}$ and $\mathcal{T}$ are more closely related to each other than they are to species $\mathcal{S}$. It also tells us that all three observed characters descend from an ancestral (unobserved) character $v \in \{A, C, G, T\}$ and that characters 'A' and 'A' descend from a common (also unobserved) ancestral character $u \in \{A, C, G, T\}$. The evolution of these characters along a prescribed tree is modelled by a continuous-time time-homogeneous Markov process on the state space $\mathcal{I} = \{A, C, G, T\}$. The Markov process models how an ancestral character is substituted by another along the branches of the tree. (We refer to the 'homogeneous phylogenetic model' as such not because of the time-homogeneity in the Markov process but because of the assumption of a common process of evolution operating on all sites.) A sample path of the substitution of characters for this example is shown in Figure 3-2(b). Two processes (corresponding to the evolution of characters at the top and bottom branches of the tree in Figure 3-2(a)) start at state C with a certain probability. Each process evolves independently and according to stochastically identical rules, following the path dictated by the given phylogenetic tree. After evolving for a length t_1, the process at the top lineage splits into two independent processes that continue to experience character substitutions for a length $t_2 - t_1$. The independent substitution mechanisms ultimately all reach their final states, which are prescribed by the observed characters at the particular site in the alignment. Then, the lineage at the bottom lands at state 'G' after evolving for a length t_2, while the top two lineages both land at state 'A'.

As suggested by this example, the tree topology is not the only parameter of interest in our studies. The lengths of the branches contain valuable evolutionary information and are quantities to be estimated too.

3.3.1 Calculating the likelihood function

The likelihood function for the phylogenetic tree in Figure 3-2(a) (denoted by ϕ), a set of branch lengths, $\boldsymbol{t} = (t_1, t_2)$, and a (possibly vector) parameter θ that specifies the generator matrix of the Markov process of nucleotide substitution, is:

$$L(\phi, \boldsymbol{t}, \theta | x_1) = \sum_{v \in \mathcal{I}} \sum_{u \in \mathcal{I}} p(v)\, p_{vu}(t_1)\, p_{vG}(t_2)\, p_{uA}(t_2 - t_1)\, p_{uA}(t_2 - t_1) \qquad (3.1)$$

where the summations are over all possible values of the ancestral characters u and v. Probability $p(v)$ is the probability that, at a random point on an evolving lineage, we would observe character $v \in \{A, C, G, T\}$. The homogeneous phylogenetic model postulates that evolution has been proceeding for a very long time so that the Markov process of nucleotide substitution has reached equilibrium at the time of the first split (or root of the tree). Therefore, $p(v)$ is taken to be the stationary probability for character v, denoted as π_v. Quantities of the form $p_{ij}(t)$ are the transition probabilities of the Markov process of nucleotide substitution (see equation (2.17)). It is the assumption of independent evolution at different lineages that allows the multiplication of the individual transition probabilities when calculating the likelihood as in (3.1).

Equation (3.1) computes the likelihood for a specific tree, the one shown in Figure 3-2(a). A different tree would require a reformulation of the likelihood according to the given branching structure. Reasonably-sized phylogenetic analyses depend on the efficient calculation of the likelihood function. An algorithm that exploits the recursive structure of the phylogenetic likelihood function is presented in Section 3.6.

Having specified the likelihood at a site, the likelihood on the joint data $\boldsymbol{x} = (x_1, \ldots, x_N)$ is the product of the site likelihoods, from site 1 to N. This is,

$$L(\phi, \boldsymbol{t}, \theta | \boldsymbol{x}) = \prod_{n=1}^{N} L(\phi, \boldsymbol{t}, \theta | x_n). \qquad (3.2)$$

It remains to specify the form of the Markov transition probabilities. From (2.19), we know that these probabilities are gathered in a transition matrix $P(t)$. The transition matrix is calculated by exponentiation of a generator matrix Q. There are very many different Q-matrices that generate a Markov process of nucleotide substitution. In particular, we concentrate on those that generate a reversible process; this is a process where the probability of starting at state i and this character being substituted by a character j over a branch of length t is the same as starting at j and evolving to i over the same length t. In other words, the phylogenetic model does not distinguish where the process started, either at i to evolve to j over t, or the other way round. This has important implications concerning the impossibility of inferring rooted phylogenetic trees – trees where one is certain that

the process started at a particular root node –, an issue that will be revisited in Section 3.5. For the moment, we will review some of the most well-known Q-matrices.

3.3.2 The JC model

The simplest model of nucleotide substitution is the one by Jukes and Cantor, known as the JC model [57]. It assumes that a change from i to j occurs at rate α for all $i, j \in \mathcal{I},\ j \neq i$. The Q-matrix is given by

$$Q(\alpha) = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \tag{3.3}$$

The rows indicate the rate of substitution from state i and the columns indicate the rate of substitution to state j. The computation of the transition probabilities by exponentiation of the Q-matrix is straightforward in this case. It requires first to factorise the rate matrix as $Q = T\Lambda T^{-1}$, where T is a matrix whose columns correspond to the eigenvectors of Q and Λ is a diagonal matrix whose entries are the eigenvalues corresponding to the columns of T. The transition probability matrix is then obtained by taking the product $T\, e^{\Lambda t}\, T^{-1}$ as follows

$$e^{Qt} = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{-4\alpha t} & 0 & 0 \\ 0 & 0 & e^{-4\alpha t} & 0 \\ 0 & 0 & 0 & e^{-4\alpha t} \end{pmatrix} \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \end{pmatrix}$$

The resulting Jukes-Cantor transition probabilities are

$$p_{ij}(t, \alpha) = \begin{cases} \frac{1}{4} + \frac{3}{4} e^{-4\alpha t} & (i = j) \\ \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} & (i \neq j) \end{cases} \tag{3.4}$$

3.3.3 The K2P model

The rules of evolution in (3.4) assume that a character i can be replaced by a character j ($j \neq i$) with equal probability. Molecular biologists know, however, that substitutions where an 'A' is replaced by a 'G' or a 'C' by a 'T' (or vice versa in each case), are more common than any of the other possible replacements. A replacement $A \leftrightarrow G$ or $C \leftrightarrow T$ is called a *transition* (the name just being a coincidence with Markov transitions) while a replacement $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ or $G \leftrightarrow T$, is called a *transversion*. The bias that

favours transitions over transversions, called the *transition bias*, can be explained in terms of the process of DNA replication.

In order for the genetic information stored in DNA to be inherited between generations it must first be replicated, so that new copies are made from parental DNA strands and passed on to the offspring. Accurate replication is crucial in the reproduction of living organisms and, although generally a very efficient and exact process, it does not function correctly on every occasion. Sometimes errors, or *mutations*, occur that cause the substitution of one character for another in the descendant DNA sequence (something called a *point mutation*). Although not all mutations are point mutations and not only DNA replication causes mutations, focusing on point mutations caused by DNA replication facilitates our discussion.

Once DNA information has been replicated, certain portions of it, called *protein-coding regions*, will serve the essential purpose of being turned into proteins. Proteins are responsible for practically every function of all living entities; compounds such as enzymes, hormones and antibodies are all proteins and they determine whether an organism lives or dies. A protein is a complex molecule made up of basic building bricks called *amino acids*. These bricks are linked together as a chain and the identity of a protein is determined, among other things, from the precise order of its component amino acids. A chain of amino acids is produced from a DNA coding-region by mapping every consecutive, non-overlapping set of three DNA characters into one amino acid. Each DNA triplet is called a *codon*, and positions within a codon are referred to as first, second and third codon positions (see Figure 3-3). The mapping from codon to amino acid is such that more than one codon may code for the same amino acid (see for example codons CTG and CTT in the top sequence in Figure 3-3; they both map to the same amino acid Leu). Consequently, some character substitutions within a codon triplet may not cause a change in the resulting amino acid, while others do. One of the suggested causes for transition bias [16] is the fact that only about 3% of transitions at the third codon position cause amino acid changes, compared with 41% of transversions. That is, an offspring whose genetic material contains mutations at the third codon position that are of the transition type ($A \leftrightarrow G$ or $C \leftrightarrow T$), has a fairly high probability of being functionally identical to its progenitor, while an offspring with substitutions at this same codon position but of the transversion type ($A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$ or $G \leftrightarrow T$) will functionally differ from its progenitor with a higher probability.

Figure 3-3 illustrates this; the top DNA sequence is the genetic material of a progenitor (with its corresponding amino acid chain shown in the uppermost line), while the bottom DNA sequence is the descendant sequence after replication (with its corresponding amino acid chain above). The replacement of a parental T by a descendant C (a transition $T \rightarrow C$) at the third codon position of codon TTT does not cause any change in the resulting amino acid of the descendant sequence. In contrast, a change from a parental

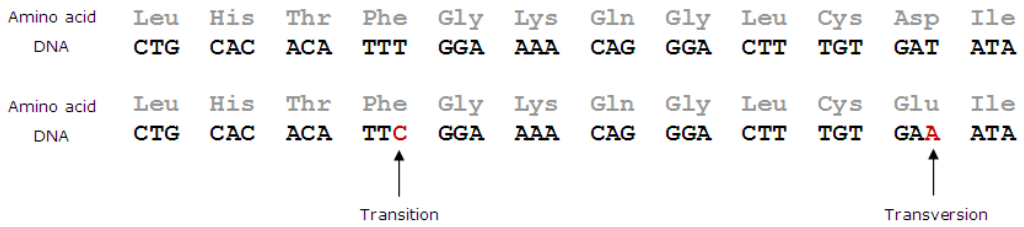


Figure 3-3: A parental DNA sequence (at the top) is replicated to give birth to a descendant DNA sequence (at the bottom), the latter shown with two point mutations in red. The point mutation to the left is of the transition type and does not cause a change in the resulting amino acid (the chain of amino acids is shown at the top of both DNA sequences, in light grey) while the mutation to the right is a transversion and causes that the descendant DNA sequence codes for amino acid Glu at that position instead of the original Asp.

T to a descendant A (a transversion $T \rightarrow A$) at the third position of codon GAT causes a change in amino acid from Asp to Glu (which is an abbreviation for Aspartic acid and Glutamic acid, respectively). Because mutations that change the amino acid can have such disastrous effects on organisms and transversions are more likely to cause a change in the amino acid chain, it is not surprising that transversions are less common than transitions [119].

A rate matrix that accounts for the different rates at which transitions and transversions occur was introduced by the geneticist Motoo Kimura in 1980 [59]:

$$Q(\boldsymbol{\theta}) = \begin{pmatrix} -\alpha - 2\beta & \beta & \alpha & \beta \\ \beta & -\alpha - 2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha - 2\beta & \beta \\ \beta & \alpha & \beta & -\alpha - 2\beta \end{pmatrix} \tag{3.5}$$

where $\boldsymbol{\theta} = (\alpha, \beta)$, and the states are ordered A, C, G, T from left to right (for example, the rate of substitution from $A \rightarrow G$ is α and the rate of substitution from $T \rightarrow G$ is β). According to this model, known as the K2P model, the probability that a character j exists after a branch of length t if the character at the start of the branch is i, is:

$$p_{ij}(t, \boldsymbol{\theta}) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t} & (i = j) \\ \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t} & (i \neq j, \text{ transition}) \\ \frac{1}{4} - \frac{1}{4}e^{-4\beta t} & (i \neq j, \text{ transversion}) \end{cases} \tag{3.6}$$

If a JC or a K2P substitution process is run over a very long branch length, i.e. if $t \rightarrow \infty$, the transition probabilities converge to the stationary probability distribution $\pi = (\pi_A = \frac{1}{4}, \pi_C = \frac{1}{4}, \pi_G = \frac{1}{4}, \pi_T = \frac{1}{4})$ (see equations (3.4) and (3.6)). The implications of

this is that a DNA sequence that is observed at a leaf node of the tree is expected to have the same number of A's, C's, G's, and T's. Such a perfect balance is not observed in nature. For instance, the sequence of human mitochondrial DNA that we will analyse in Chapter 7 contains nucleotides in the following proportions:

A	C	G	T
0.304	0.331	0.109	0.256

This justifies the need to relax the constraint of equal stationary probabilities by specifying a Q-matrix that accommodates different π_A, π_C, π_G and π_T.

3.3.4 The HKY85 model

A more realistic model of nucleotide substitution should account for both the different rates at which transitions and transversions occur, along with the difference in stationary probabilities. Hasegawa, Kishino and Yano [45] introduced such a model in 1985, known as the HKY85 model, with rate matrix

$$Q(\boldsymbol{\theta}) = \begin{pmatrix} -\alpha\pi_G - \beta\pi_Y & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & -\alpha\pi_T - \beta\pi_R & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & -\alpha\pi_A - \beta\pi_Y & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & -\alpha\pi_C - \beta\pi_R \end{pmatrix} \tag{3.7}$$

where $\boldsymbol{\theta} = (\alpha, \beta, \boldsymbol{\pi})$, $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$, $\pi_Y = \pi_C + \pi_T$, and $\pi_R = \pi_A + \pi_G$. (As before, the states are ordered A, C, G, T from left to right; thus, the rate of substitution from $A \to G$ is $\alpha\pi_G$ while the rate of substitution from $T \to G$ is $\beta\pi_G$.) The HKY85 Q-matrix is non-symmetric and, in order to compute the transition probabilities, it is necessary to first find a symmetric matrix that is related to the non-symmetric Q and that leads to the easy computation of the eigenvalues and eigenvectors of Q. Details of the exponentiation of a non-symmetric rate matrix are presented in Section 3.4.1.

3.3.5 The GTR model

There exists an even more general time-reversible model, known as the GTR model, that allows all nucleotides to be substituted at different rates while incorporating different stationary probabilities. It was first described by Tavaré in 1986 [115], and is expressed as a matrix of the form

$$Q(\boldsymbol{\theta}) = \begin{pmatrix} q_{AA} & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & q_{CC} & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & q_{GG} & r_{GT}\pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & q_{TT} \end{pmatrix} \tag{3.8}$$

where $\boldsymbol{\theta} = (\boldsymbol{r}, \boldsymbol{\pi})$, and the diagonal elements in Q are defined as $q_{ii} = -\sum_{j \in \mathcal{I}; j \neq i} q_{ij}$. Vector $\boldsymbol{r}$ contains six *substitution rates*, $\boldsymbol{r} = (r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT})$ and $\boldsymbol{\pi} = (\pi_i,\, i \in \mathcal{I})$ is the *stationary distribution* of the Markov process. The substitution rates in $\boldsymbol{r}$ are positive real quantities constrained to sum to one.

According to Huelsenbeck, Larget and Alfaro [50], there are over 203 substitution models that satisfy time reversibility, all being special cases of the GTR model. For example, the JC model is obtained by setting $\pi_A = \ldots = \pi_T = \frac{1}{4}$ and $r_{AC} = \ldots = r_{GT} = \lambda$ in matrix (3.8); this results in $\alpha = \frac{\lambda}{4}$ in the Jukes-Cantor matrix (3.3). If the stationary probabilities remain equal and the substitution rates are set to $r_{AG} = r_{CT} = \lambda$ and $r_{AC} = r_{AT} = r_{CG} = r_{GT} = \kappa$, matrix (3.8) becomes the K2P matrix in (3.5), with $\alpha = \frac{\lambda}{4}$ and $\beta = \frac{\kappa}{4}$. The GTR model is our preferred description of the process of nucleotide substitution since it is the most general of the time-reversible formulations. In the following, the reader should assume that the Q-matrix of the Markov process of substitution refers to the GTR model unless otherwise specified.

3.4 Algebra on Q-matrices

3.4.1 Exponentiation of a non-symmetric matrix

The calculation of the transition probability matrix involves the exponentiation of the Q-matrix, i.e. e^{Qt}, which in turn requires the computation of the eigenvalues and eigenvectors of Q. This is relatively straightforward whenever Q is symmetric, since a symmetric matrix has only real eigenvalues and eigenvectors. A non-symmetric Q-matrix, however, may have complex eigenvalues and eigenvectors, and computing those eigenvalues and eigenvectors is much harder. In this case, additional intermediate steps are necessary in order to calculate the transition probability matrix. The idea is to find a symmetric matrix that is related to the non-symmetric Q in some way, and that leads to the easy computation of the eigenvalues and eigenvectors of this non-symmetric Q. The typical approach is to rewrite the non-symmetric Q as the product of two symmetric matrices. For example, the HKY85 rate matrix in (3.7) can be rewritten as:

$$Q = \begin{pmatrix} -W & \beta & \alpha & \beta \\ \beta & -X & \beta & \alpha \\ \alpha & \beta & -Y & \beta \\ \beta & \alpha & \beta & -Z \end{pmatrix} \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix} \tag{3.9}$$

where the dependence of Q on $\boldsymbol{\theta}$ has been dropped for simplicity, and W, X, Y and Z can be found from the diagonal entries in matrix (3.7). For instance, $W = \frac{\alpha\pi_G + \beta\pi_Y}{\pi_A}$. Denote by B and D the two matrices on the right-hand side of (3.9) and compute a matrix of the form $S := D^{\frac{1}{2}} Q D^{-\frac{1}{2}}$. It can easily be shown that S is symmetric by taking its transpose S^T:

$$\begin{aligned} S^T &= (D^{\frac{1}{2}}\, Q\, D^{-\frac{1}{2}})^T \\ &= (D^{\frac{1}{2}}\, BD\, D^{-\frac{1}{2}})^T \qquad \text{(since } Q := BD\text{)} \\ &= (D^{\frac{1}{2}}\, B \quad {}^{\frac{1}{2}})^T \\ &= (D^{\frac{1}{2}})^T\, B^T\, (D^{\frac{1}{2}})^T \\ &= D^{\frac{1}{2}}\, B\, D^{\frac{1}{2}} \qquad \text{(since } B^T = B \text{ and } (D^{\frac{1}{2}})^T = D^{\frac{1}{2}} \text{ as symmetric)} \\ &= D^{\frac{1}{2}}\, BD\, D^{-\frac{1}{2}} \\ &= S \end{aligned}$$

Hence, S is symmetric. For the HKY85 case, this matrix looks like:

$$S = \begin{pmatrix} -\alpha\pi_G - \beta(\pi_C + \pi_T) & \beta\sqrt{\pi_A\pi_C} & \alpha\sqrt{\pi_A\pi_G} & \beta\sqrt{\pi_A\pi_T} \\ \beta\sqrt{\pi_A\pi_C} & -\alpha\pi_T - \beta(\pi_A + \pi_G) & \beta\sqrt{\pi_C\pi_G} & \alpha\sqrt{\pi_C\pi_T} \\ \alpha\sqrt{\pi_A\pi_G} & \beta\sqrt{\pi_C\pi_G} & -\alpha\pi_A - \beta(\pi_C + \pi_T) & \beta\sqrt{\pi_G\pi_T} \\ \beta\sqrt{\pi_A\pi_T} & \alpha\sqrt{\pi_C\pi_T} & \beta\sqrt{\pi_G\pi_T} & -\alpha\pi_C - \beta(\pi_A + \pi_G) \end{pmatrix}$$

Any real symmetric matrix is diagonalisable (see for example [106, ch. 5]) and diagonalisation of S can be accomplished by $\Delta = U^{-1}SU$, where Δ is a diagonal matrix. The decomposition

$$U\Delta U^{-1} = S \tag{3.10}$$

thus corresponds to a matrix U whose columns are the eigenvectors of S and a diagonal matrix Δ of the eigenvalues corresponding to the columns of U. Since $S := D^{\frac{1}{2}} Q D^{-\frac{1}{2}}$, we have that $D^{-\frac{1}{2}} S D^{\frac{1}{2}} = Q$ and

$$
\begin{aligned}
Q^n &= (D^{-\frac{1}{2}}\, S\, D^{\frac{1}{2}})^n \\
&= D^{-\frac{1}{2}}\, S^n\, D^{\frac{1}{2}} \\
&= D^{-\frac{1}{2}}\, (U \Delta U^{-1})^n\, D^{\frac{1}{2}} \qquad \text{(from factorisation (3.10))} \\
&= D^{-\frac{1}{2}}\, U \Delta^n U^{-1}\, D^{\frac{1}{2}} \\
&= C\, \Delta^n\, C^{-1} \qquad \text{(by letting } C := D^{-\frac{1}{2}}\, U\text{, an invertible matrix)}
\end{aligned}
\tag{3.11}
$$

where Δ^n is a diagonal matrix and can be trivially computed by raising each entry to the nth power. This shows how the calculation of Q^n is greatly simplified once having factorised S as in (3.10). Finally, the matrix of transition probabilities is easily obtained as:

$$
\begin{aligned}
P(t) &= e^{Qt} \\
&= \sum_{n=0}^{\infty} \frac{Q^n\, t^n}{n!} \\
&= \sum_{n=0}^{\infty} \frac{C\, \Delta^n\, C^{-1}\, t^n}{n!} \qquad \text{(from factorisation (3.11))} \\
&= C \left(\sum_{n=0}^{\infty} \frac{\Delta^n\, t^n}{n!} \right) C^{-1} \\
&= C\, e^{\Delta\, t}\, C^{-1}
\end{aligned}
$$

and, since Δ is a diagonal matrix of the form $\Delta = diag\{a_1, a_2, \ldots, a_k\}$ say, its exponential can be easily computed as $e^{\Delta t} = diag\{e^{a_1 t}, e^{a_2 t}, \ldots, e^{a_k t}\}$. The implementations associated with this research have frequently dealt with the exponentiation of rate matrices; in doing so, the method above presented has proved essential. A further discussion of the exponentiation of non-symmetric rate matrices within a phylogenetic context can be found in [27, ch. 16].

3.4.2 Standardisation of the rate matrix

The formulation of the homogeneous phylogenetic model is such that branch lengths are the product of rate and time. In equation (3.4), for example, the probability of transition from character i to j depends on t only through the product αt. In this equation, if we were to double α and halve t, there would be no change in the resulting p_{ij}. Under this formulation, all we can infer is the product of rate and time, αt, and not α or t individually.

The discussion so far has focused upon a particular scenario in which all entities across the tree obey the same Markov process of nucleotide substitution. If the rate matrix is the same for all lineages and the process is run for a given period of time, then the resulting tree has leaf nodes that align at the same distance from the root (just like the trees in

Figure 3-1). This over-simplistic scenario was only used for illustrative purposes and does not fully correspond to the type of trees with which this study is concerned.

A biologically valid model of the evolution of organisms should accommodate different rate matrices operating on different branches of the tree. The reason for this is the different biology that distinct entities in the tree may be obeying (e.g. distinct generation times, variation in population sizes, or unequal environmental conditions). Suppose that two species $\mathcal{R}$ and $\mathcal{S}$, having descended from a hypothetical common ancestor $\mathcal{Z}$, experience substitutions according to rate matrices Q_1 and Q_2 respectively. If branch lengths are the result of rate and time acting together, it is far more convenient to assume that processes along all branches are generated from the same rate matrix, and to vary the time of evolution instead.

In order to illustrate this, let the *expected total rate of substitution* of a Markov substitution process with generator matrix $Q = (q_{ij},\, i, j \in \mathcal{I})$ be given by

$$\kappa = \sum_{i \in \mathcal{I}} -q_{ii}\pi_i. \tag{3.12}$$

(This, since the total rate at which the process tries to leave state i is $-q_{ii}$ and the stationary process of substitution is at state i with probability π_i.) Suppose that species $\mathcal{R}$ and $\mathcal{S}$ evolve for a period of time $t \geq 0$ under substitution processes generated by Q_1 and Q_2 respectively, with corresponding expected total rates κ_1 and κ_2 ($\kappa_1, \kappa_2 > 0$, $\kappa_1 \neq \kappa_2$). Then, the branch leading to species $\mathcal{R}$ has an expected length $\kappa_1 t$, while the branch that leads to $\mathcal{S}$ is expected to be $\kappa_2 t$ units long. Notice how these two branches, being the product of expected total rate and time, represent the expected number of nucleotide substitutions or *expected amount of evolution* between the two connected nodes. In the absence of external evidence about the expected total rate of substitution, Felsenstein [23] suggested adopting the convention that $\kappa_1 = \kappa_2 = 1$ and varying the time of evolution at each branch accordingly, so that the expected number of substitutions remains unchanged. With κ_1 and κ_2 fixed to one, we can re-express the expected branch lengths of $\mathcal{R}$ and $\mathcal{S}$ as the product of the κs with a branch-specific period of time as $\kappa_1\, t_1 = t_1$ and $\kappa_2\, t_2 = t_2$, respectively. A branch of length t_1 is therefore the expected amount of evolution and not time (one unit of branch length is the segment in which we expect to see one nucleotide substitution). This is the interpretation that we adopt for the branch lengths in this thesis. The reader might find it convenient, however, to think of them as the period of 'time' during which the Markov process of substitution operates, but one must keep in mind what branch lengths actually represent.

An example of the standardisation of the Jukes-Cantor rate matrix in (3.3), so that the expected total rate of substitution is one follows. The non-standard rate matrix Q needs to be scaled by a standardising factor μ, which results in

$$Q\mu = \begin{pmatrix} -3\alpha\mu & \alpha\mu & \alpha\mu & \alpha\mu \\ \alpha\mu & -3\alpha\mu & \alpha\mu & \alpha\mu \\ \alpha\mu & \alpha\mu & -3\alpha\mu & \alpha\mu \\ \alpha\mu & \alpha\mu & \alpha\mu & -3\alpha\mu \end{pmatrix} \tag{3.13}$$

Using condition $\kappa = 1$ (see equation (3.12)) to solve for μ, we obtain

$$\begin{aligned} \sum_{i\in\mathcal{I}} (3\alpha\mu)\,\pi_i &= 1 \\ (3\alpha\mu) \sum_{i\in\mathcal{I}} \pi_i &= 1 \\ \mu &= \frac{1}{3\alpha} \end{aligned}$$

The standardised Jukes-Cantor rate matrix is finally given by

$$Q\mu = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} \tag{3.14}$$

One of the consequences of allowing rates of substitution to differ from segment to segment on the tree is that the leaf nodes are not aligned at an equal distance from the root anymore. This produces more realistic scenarios of evolution and it will be the approach taken in this thesis.

3.5 On the impossibility of inferring rooted trees

The homogeneous phylogenetic model postulates that the Markov process of nucleotide substitution is time-reversible. That is, the probability of starting at state i and ending at state j after a segment t is the same as the probability of starting at j and evolving to i over the same t,

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t) \tag{3.15}$$

for $i, j \in \mathcal{I}$ and $t \geq 0$. The reversibility assumption has an important implication for the estimation of trees, namely that all we can infer under this model is an unrooted and not a rooted tree.

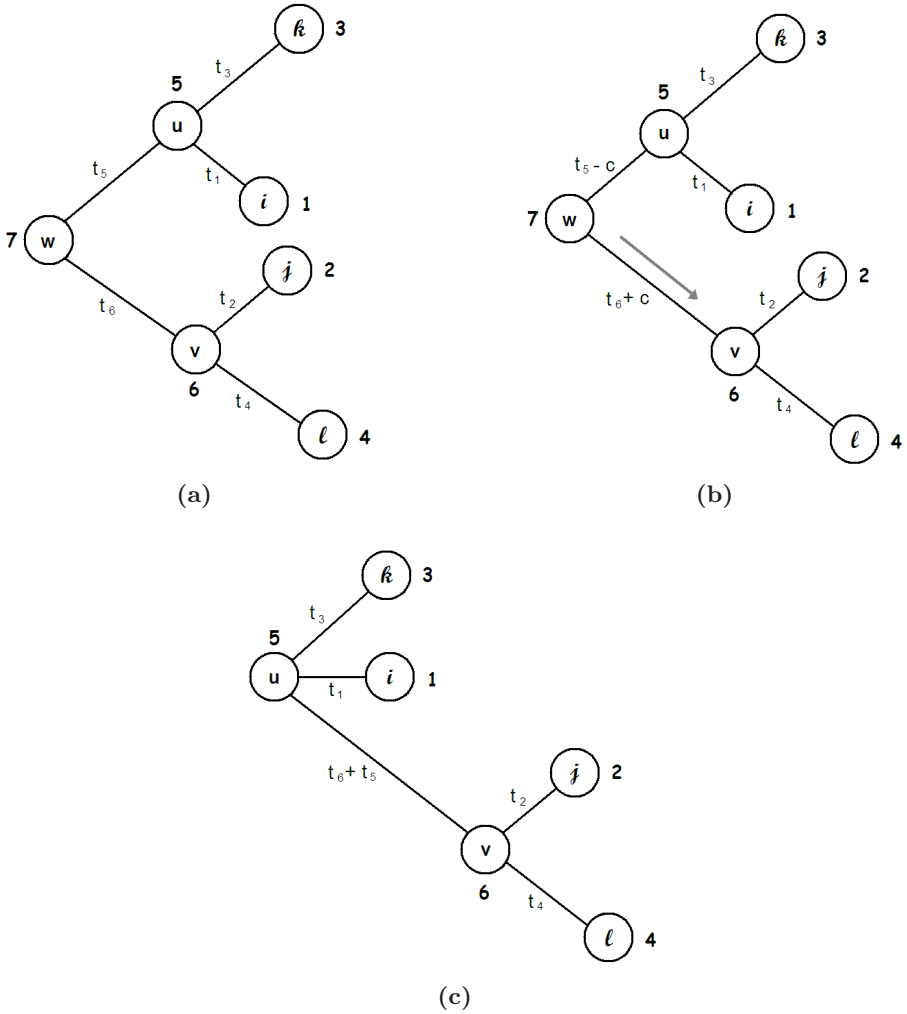


Figure 3-4: These trees illustrate the impossibility of inferring rooted trees with the homogeneous phylogenetic model. In terms of the probability of x_n, none of the trees in this figure can be distinguished from one another.

Consider the phylogenetic tree, denoted by ϕ, with set of branch lengths $\boldsymbol{t} = (t_1, \ldots, t_6)$ shown in Figure 3-4(a). Suppose that a Markov process generated by a Q-matrix that is parametrised by $\boldsymbol{\theta}$ is used to describe the substitution of characters along the branches of this tree. Given a DNA alignment of four sequences, the probability of observing characters $x_n = (i, j, k, l)^T$ at site n which evolve under ϕ, $\boldsymbol{t}$ and $\boldsymbol{\theta}$ is

$$p(x_n|\phi, \boldsymbol{t}, \boldsymbol{\theta}) = \sum_{u\in\mathcal{I}} \sum_{v\in\mathcal{I}} \sum_{w\in\mathcal{I}} \pi_w p_{wu}(t_5) p_{wv}(t_6) p_{ui}(t_1) p_{uk}(t_3) p_{vj}(t_2) p_{vl}(t_4).$$

By invoking the time-reversibility property (3.15) and the Chapman-Kolmogorov equation (2.18), it is possible to rewrite this as

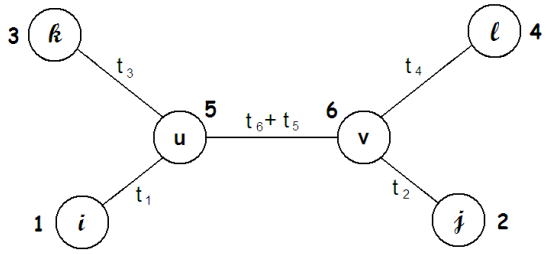


Figure 3-5: The type of unrooted tree that is possible to estimate under the homogeneous phylogenetic model.

$$
\begin{aligned}
p(x_n|\phi,\boldsymbol{t},\boldsymbol{\theta}) &= \sum_{u\in\mathcal{I}}\sum_{v\in\mathcal{I}}\sum_{w\in\mathcal{I}} \underbrace{\pi_u p_{uw}(t_5)}_{\text{since } \pi_w p_{wu}(t_5)=\pi_u p_{uw}(t_5)} p_{wv}(t_6)p_{ui}(t_1)p_{uk}(t_3)p_{vj}(t_2)p_{vl}(t_4) \\
&= \sum_{u\in\mathcal{I}}\sum_{v\in\mathcal{I}}\pi_u \sum_{w\in\mathcal{I}} p_{uw}(t_5)p_{wv}(t_6)\; p_{ui}(t_1)p_{uk}(t_3)p_{vj}(t_2)p_{vl}(t_4) \\
&= \sum_{u\in\mathcal{I}}\sum_{v\in\mathcal{I}}\pi_u \underbrace{p_{uv}(t_5+t_6)}_{\text{since } \sum_{w\in\mathcal{I}} p_{uw}(t_5)p_{wv}(t_6)=p_{uv}(t_5+t_6)} p_{ui}(t_1)p_{uk}(t_3)p_{vj}(t_2)p_{vl}(t_4)
\end{aligned}
\tag{3.16}
$$

This derivation shows how the probability is unaffected if we add a length $c \geq 0$ to t_6 and substract the same length from t_5 (see Figure 3-4(b)). In fact, tree ϕ could have its root anywhere between nodes 5 and 6, and the probability would not change. Felsenstein [23] dubbed this the *pulley principle* since the root acts as a pulley. If all parts of the tree to one side of the root are pulled away from it, and all parts to the other side are moved towards the root by the same length, the probability remains unaltered. In terms of the probability of x_n, none of the trees in Figure 3-4 can be distinguished from one another.

This argument can be applied repeatedly to show that there is no information whatsoever about the placement of the root. By applying the time-reversibility property to the last line in (3.16), where the root of the tree is node 5, we obtain

$$
p(x_n|\phi,\boldsymbol{t},\boldsymbol{\theta}) = \sum_{u\in\mathcal{I}}\sum_{v\in\mathcal{I}} \underbrace{\pi_v p_{vu}(t_5+t_6)}_{\text{since } \pi_u p_{uv}(t_5+t_6)=\pi_v p_{vu}(t_5+t_6)} p_{ui}(t_1)p_{uk}(t_3)p_{vj}(t_2)p_{vl}(t_4)
$$

which tells us once again that, in terms of the probability of x_n, the tree ϕ rooted at node 5 is indistinguishable from this same tree rooted at node 6. In fact, the root can be placed anywhere in the tree without affecting the probability of x_n. Under this model, what we are in effect estimating is an unrooted tree, such as the one in Figure 3-5, and not a rooted one.

Nevertheless, rooted trees are biologically relevant as they represent how a set of organisms evolved from a common hypothetical ancestor. There are two methods in the literature to determine the position of the root in an unrooted tree; midpoint rooting [20] and rooting by outgroup. The former places the root of the tree in the middle between the most distantly related taxa; that is, the taxa that lie the farthest apart in the unrooted tree. In Figure 3-5, the method of midpoint rooting would place the root in the middle of the path between nodes 3 and 4. The alternative is to include an *outgroup* sequence in the analysis. The outgroup sequence should contain the genetic information of an organism that is *a priori* known to be less closely related to any of the other organisms in the study (known as the *ingroup*) than any pair of ingroup organisms are related to each other. An example outgroup would be to include the sequence of a bird when analysing the phylogeny of mammals. The outgroup is treated exactly as all other organisms. When a final (unrooted) tree is estimated, the root is placed at any point on the branch that connects the ingroup organisms to the outgroup.

3.6 Felsenstein's pruning algorithm

Suppose that $x_n = (G, T, A, A)^T$ is the set of observed nucleotides at the nth site of a DNA alignment of four sequences. The probability of observing those characters given the phylogenetic tree and the set of branch lengths in Figure 3-5, and a vector of Q-matrix parameters $\boldsymbol{\theta}$, is computed by arbitrarily rooting the tree at any node (Section 3.3.1). Suppose that node 5 is the root, this causes the tree to look like the one in Figure 3-6. The likelihood is then calculated as

$$L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|x_n) = \sum_{u \in I} \sum_{v \in I} \pi_u p_{uG}(t_1) p_{uA}(t_3) p_{uv}(t_6 + t_5) p_{vT}(t_2) p_{vA}(t_4) \tag{3.17}$$

This expression has 16 terms and, in general, the probability for S species will have 4^{S-2} terms. The number of terms can be very large and calculating this probability may be computationally prohibitive. A recursive technique for its efficient computation, called the *pruning algorithm*, was introduced by Joseph Felsenstein in 1981 [22, 23]. By moving the summation signs as far right as possible, one gets a flow of computation that corresponds to starting at the leaf nodes and moving toward the root. The algorithm restates the calculation in terms of probabilities of a subtree as follows.

Definition 3.6.1. *Let $L_r^{(n)}(m)$ be the probability of everything that is observed at or 'below' node r on a (arbitrarily) rooted phylogenetic tree given that node r is known to have state m for site n.*

If node r is a leaf, $L_r^{(n)}(m)$ is zero for all states except for the state actually observed. For node 1 in Figure 3-6, the vector of these probabilities is

$$(L_1^{(n)}(A),\, L_1^{(n)}(C),\, L_1^{(n)}(G),\, L_1^{(n)}(T)) = (0,\, 0,\, 1,\, 0).$$

For node 6, whose children nodes are 2 and 4, the probability is

$$L_6^{(n)}(m) = \underbrace{\left(\sum_{y\in\mathcal{I}} p_{my}(t_2)L_2^{(n)}(y)\right)}_{\text{(all terms are } 0 \text{ except when } y=T)} \overbrace{\left(\sum_{z\in\mathcal{I}} p_{mz}(t_4)L_4^{(n)}(z)\right)}^{\text{(all terms are } 0 \text{ except when } z=A)}.$$

Likewise, the probability of the subtrees at or 'below' node 5 is

$$L_5^{(n)}(m) = \left(\sum_{y\in\mathcal{I}} p_{my}(t_1)L_1^{(n)}(y)\right)\left(\sum_{z\in\mathcal{I}} p_{mz}(t_3)L_3^{(n)}(z)\right)\left(\sum_{w\in\mathcal{I}} p_{mw}(t_6+t_5)L_6^{(n)}(w)\right).$$

It is then possible to rewrite the likelihood in (3.17) as

$$L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|x_n) = \sum_{u\in\mathcal{I}} \pi_u L_5^{(n)}(u).$$

As before, by the site-independence assumption, the likelihood function on the joint data $\boldsymbol{x} = (x_1, \ldots, x_N)$ is finally computed as

$$L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|\boldsymbol{x}) = \prod_{n=1}^{N} L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|x_n). \tag{3.18}$$

The pruning algorithm plays an essential role in the implementations carried out during this research.

3.7 The model and choice of priors

A Bayesian approach to the inference of phylogenetic parameters starts by providing a joint probability distribution for the parameters $\phi, \boldsymbol{t}, \boldsymbol{\theta}$ and the data $\boldsymbol{x}$ as follows

$$\begin{aligned} p(\phi, \boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{x}) &= p(\phi, \boldsymbol{t}, \boldsymbol{\theta}) p(\boldsymbol{x}|\phi, \boldsymbol{t}, \boldsymbol{\theta}) \\ &= p(\phi) p(\boldsymbol{t}) p(\boldsymbol{\theta}) L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|\boldsymbol{x}) \end{aligned} \tag{3.19}$$

In this model, all phylogenetic trees on the label set $(1, 2, \ldots, S)$ are taken to be equally likely *a priori*, so that

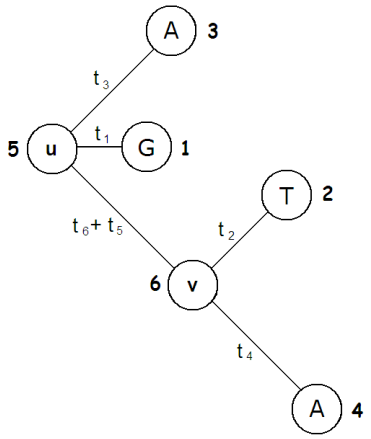


Figure 3-6: A phylogenetic tree on the label set $(1, \ldots, 4)$ with set of branch lengths $\boldsymbol{t} = (t_1, t_2, \ldots, t_5 + t_6)$ used to illustrate Felsenstein's 'pruning algorithm'. This tree has been arbitrarily rooted at node 5 as the homogeneous phylogenetic model does not distinguish between alternative rootings.

$$p(\phi = \phi_r) = \frac{1}{|\Phi(S)|}, \quad r = 1, \ldots, |\Phi(S)| \tag{3.20}$$

where $|\Phi(S)|$ is the number of all binary unrooted phylogenetic trees with labels $(1, 2, \ldots, S)$ (Section 2.8.2). Vector $\boldsymbol{t}$ contains the $2S-3$ individual branch lengths over which we specify priors of the form

$$t_h \sim Exp(\beta), \quad \text{independently for } h = 1, \ldots, 2S-3 \tag{3.21}$$

where $\beta > 0$ is the rate parameter of the exponential distribution with $\mathbb{E}(t_h) = \frac{1}{\beta}$. This choice of prior obeys the formulation of the substitution of characters as a Markov process. It is known, from Markov theory, that a continuous-time time-homogeneous Markov process remains in a particular state for an exponentially distributed amount of time before making a jump to a new state (see [44, ch. 6] for example). In the phylogenetic model, the length of a branch is related to the amount of 'time' during which the process evolves, hence the convenience of assigning an exponential prior to individual branch lengths (e.g. [112, 88, 89]).

The evolutionary parameters in $\boldsymbol{\theta}$ include a vector of six substitution rates, $\boldsymbol{r} = (r_{AC}, \ldots, r_{GT})$, and a vector of four stationary probabilities $\boldsymbol{\pi} = (\pi_A, \ldots, \pi_T)$. The model defines independent prior distributions for $\boldsymbol{r}$ and $\boldsymbol{\pi}$ of the form:

$$\begin{aligned} \boldsymbol{r} &\sim Dir_6(1, \ldots, 1) \\ \boldsymbol{\pi} &\sim Dir_4(1, \ldots, 1) \end{aligned} \tag{3.22}$$

where the probability density function of a random variable $(y_1, \ldots, y_g) \sim Dir_g(\alpha_1, \ldots, \alpha_g)$ is given by

$$f(\boldsymbol{y}) = \frac{\Gamma(\sum_{i=1}^{g} \alpha_i)}{\prod_{i=1}^{g} \Gamma(\alpha_i)} \prod_{i=1}^{g} y_i^{\alpha_i - 1} \tag{3.23}$$

and $\Gamma(a)$ is the gamma function $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$. After specifying the prior distributions for model parameters it is possible to rewrite the joint probability (3.19) as

$$p(\phi, \boldsymbol{t}, \boldsymbol{\theta}, x) \propto e^{-\beta \sum_{h=1}^{2S-3} t_h} \prod_{n=1}^{N} L(\phi, \boldsymbol{t}, \boldsymbol{\theta} | x_n). \tag{3.24}$$

3.8 Discussion

We have presented the conventional probabilistic model for phylogenetic observations. In this model, a number of simplifying assumptions are required to make the inference of parameters mathematically tractable. Unfortunately the price of tractability is that the model may lose realism in scenarios where the assumptions do not closely correspond to reality. The assumption of 'independence of evolution among sites', for example, prevents the homogeneous model from adequately describing the evolution of RNA (ribonucleic acid). RNA is a molecule that, as DNA, carries genetic information. This molecule plays a critical role in the production of proteins and works as a link between the nucleus and the rest of the cell. A characteristic of RNA is that it is a single-stranded molecule and, as such, has a natural urgency to pair and turn itself into a double-stranded molecule. The way in which RNA forms pairs has a complicated structure comprising a *stem* and a *loop* (Figure 3-7). Nucleotides at the stem form pairs according to a strict complementary relationship, A only pairs with U and C only pairs with G (in RNA, there are Us instead of the usual Ts found in DNA). Nucleotides at the loop do not pair together. The substitution of a nucleotide in the stem may result in a pair of nucleotides that cannot hold together correctly anymore, reducing the stability of the molecule. For example, in Figure 3-7, a nucleotide G is substituted by a C, leaving a set of nucleotides $C\ C$ that cannot form a pair anymore. A *compensatory change* may then occur at the other strand to restore the pairing. So, in Figure 3-7, nucleotide C in the left-hand strand is changed into G to stabilise the molecule again. In a case like this, the evolution of sites that are separated many positions apart may affect one another and the assumption of independence among sites of the homogeneous model is invalid. This is further explained in the textbook by Page and Holmes [87, ch. 5], and a model that accounts for this was introduced by Gessell and Haeseler [34].

A second assumption of the homogeneous phylogenetic model is 'independence of evolution at different lineages'. This supposition can be biologically unrealistic when two

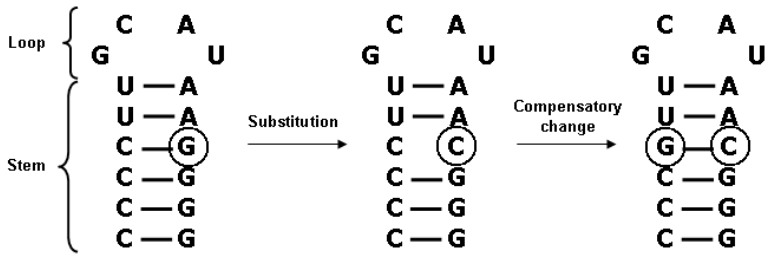


Figure 3-7: RNA molecules tend to turn themselves into double-stranded molecules by adopting a complicated structure comprising a stem and a loop. The nucleotides at the stem form pairs, and a substitution in one strand of the stem has to be compensated at the other strand to restore the pairing. The way RNA evolves contradicts the assumption of 'independence of evolution among sites' of the homogeneous phylogenetic model. After Hickson et al. (see [87, p. 157]).

lineages (representing two different species) compete for resources. The evolution of one should then affect the evolution of the other. Also, environmental effects such as changes in weather or changes in common predators or prey may affect several lineages at once and cause correlated evolution [24].

Finally, a third assumption of the homogeneous phylogenetic model is that 'a single tree, set of branch lengths and Q-matrix' are sufficient to characterise the evolutionary process across the entire DNA alignment. In scenarios where different sites in a DNA alignment accumulate a different number of substitutions, for example, the characterisation of all sites with the same set of branch lengths is inappropriate. The inferred branch lengths are a compromise among signals coming from differently evolving sites; sites accumulating fewer substitution are generated by a phylogenetic tree with short branches whereas sites experiencing a higher number of substitutions arise from a tree with long branches. A model that includes only one set of branch lengths misses both signals and recovers spurious lengths (Section 1.1). Similarly, when different sites experience substitutions according to different rules of evolution, a formulation that includes only one Q-matrix fails to capture the heterogeneity that underlies the data.

As suggested by these examples, the homogeneous model is a rather restrictive formulation. Most of this thesis will be devoted to introduce an alternative model that does not assume that 'a tree topology with a single set of branch lengths and Q-matrix' is sufficient to characterise the entire DNA alignment.

Chapter 4

MCMC methods for the homogeneous phylogenetic model

4.1 Introduction

Bayesian phylogenetic modelling poses a challenge for statistical inference due to the need to integrate over multi-dimensional parameter spaces. It is no surprise that this approach to modelling made little progress until the introduction of MCMC methods in the late 1990s. Today, although MCMC is at the heart of Bayesian phylogenetic practice, there are still a large number of aspects that remain to be investigated. These range from the systematic assessment of tree proposals to the construction of algorithms to represent the tree in a computer-readable format and methods to keep count of visited trees. In this chapter, we address these issues and describe the MCMC sampler that we have designed for estimating the parameters of the homogeneous phylogenetic model.

This chapter begins by giving a brief account of MCMC methods for phylogenetic inference and goes on to describe two of the existing mechanisms for generating candidate trees and branch lengths. One of them, usually referred to as LOCAL, is widely used in phylogenetic software. Here we discuss its poor performance and introduce a different scheme for updating tree and branch-lengths that achieves greater efficiency. We also present a novel mechanism for updating the parameters of the Q-matrix that improves the mixing of the chain (relative to proposals commonly used in the phylogenetics literature) at no extra computational cost.

4.2 A brief history of MCMC for phylogenetics

MCMC methods for phylogenetic inference were independently introduced by three research groups in the late 1990s: Yang and Rannala [131]; Mau and Newton [75]; and Li, Pearl and Doss [65]. The computational power at the time, together with the just born phylogenetic MCMC algorithms, would only allow the analysis of a modest number of DNA sequences.

Nevertheless, MCMC had already proved to be one of the most powerful methodologies for the evaluation of posterior expectations in situations in which all other analytical and numerical techniques were intractable. It was only a matter of time and computational power before the Bayesian inference of phylogenies by MCMC began to grow increasingly popular.

Rannala and Yang [93], in 1996, attempted one of the first Bayesian approaches to phylogenetic inference. In their work, they only performed inference on the phylogenetic tree and so all other parameters were estimated by frequentist techniques prior to approximating the posterior for trees. The model they used differs from the homogeneous phylogenetic model in that they specified a birth-death process as the prior for trees and branch lengths. In this first study, Rannala and Yang did not use MCMC and instead used numerical integration, which turned out to be cumbersome and impractical for more than five DNA sequences. One year later they expanded their work by introducing MCMC simulation [131]. Their MCMC sampler moved in the discrete space of phylogenetic trees according to a simple transition mechanism, however, their strategy required Monte Carlo integration at every iteration to integrate out the branch lengths, which made it computationally intensive even for a moderate number of DNA sequences.

Mau and Newton [75] constructed a Metropolis-Hastings sampler to estimate phylogenetic trees and other model parameters. They restricted attention to those trees where all lineages are assumed to evolve at equal rates, which results in rooted trees with all their leaves lying at the same distance from the root, called *clock-like trees* (Section 2.8.1). Further to studying clock-like trees, Mau and Newton also considered trees where the order of the interior nodes with respect to the root is recorded, called *labelled histories*. Dealing with clock-like labelled histories requires rooted rather than unrooted trees and there are $2S-3$ times more rooted trees than unrooted ones (where S is the number of analysed DNA sequences). Moreover, different orders of interior nodes for a given rooted tree generate different labelled histories, which means that the space of labelled histories is even larger than the space of rooted trees (which is already larger than the space of unrooted ones). This highlights the high computational cost of moving in the space of labelled histories, or even in the space of rooted trees, as compared with the space of unrooted ones. In any case, inferring labelled histories is biologically unrealistic because imposing equal rates of evolution throughout the entire tree presumes that the conditions of evolution are the same for all organisms (Section 3.4.2). In spite of this, the work by Mau and colleagues is important because they pioneered the design of MCMC proposals for generating candidate trees (albeit unnecessarily complicated ones).

It was Larget and Simon [63] who relaxed the assumption of clock-like trees and, based on Mau and Newton's work, designed proposals which equally dealt with clock-like (rooted) and non-clock-like (either rooted or unrooted) trees. Their work retains an important im-

pact on the implementation of MCMC algorithms for Bayesian phylogenetics. One of the mechanisms they designed, called *LOCAL*, is widely used in phylogenetic packages, but its block-updates of branch lengths and tree can cause bad mixing of the chain in certain applications that we describe in more detail below.

In [65], Li, Tanimura and Sharp took a slightly different modelling approach; they made inference on ancestral DNA sequences, rooted (clock-like) trees and branch lengths. To propose a transition in the space of these parameters, they sequentially modified the tree in a small neighbourhood, proposed new branch lengths and generated a candidate ancestral DNA sequence for a specific interior node, all in a single MCMC step. Their strategy suffers from the same problem as that of Larget and Simon [63]; the generation *en bloc* of a candidate tree, branch lengths and in this case even an ancestral DNA sequence, can have detrimental effects on the mixing of the chain.

As history itself suggests, a major challenge during the implementation of MCMC samplers for phylogenetic inference is the design of efficient proposal mechanisms. In particular, mechanisms for updating trees greatly differ from conventional MCMC strategies because a phylogenetic tree is a graph structure. The next section discusses the details of some of the existing algorithms for updating tree and branch lengths, which will be later used for comparison with our own proposals.

4.3 Existing tree and branch-length proposals

4.3.1 Mau and Newton's proposal

Mau and Newton [75] designed a mechanism for block-updating clock-like labelled histories and coalescence lengths[1]. Given a labelled history, the mechanism first produces a random left-right switch of the descendant nodes for each internal node (see the two left-most trees in Figure 4-1). Next, the coalescence length for each pair of leaf nodes is calculated and independently modified by drawing from a uniform distribution centered at the current length and with support over an interval of width 2δ. Quantity $\delta > 0$ is a tuning parameter that modulates the step-size of the proposal. Candidate lengths are restricted to being greater than zero and any negative proposed length is reflected back at the zero boundary.

Figure 4-1 illustrates this mechanism. On the third tree from left to right, the coalescence lengths are shown in grey. The red and blue lines indicate proposed additions/subtractions to the original lengths under two different scenarios. Let us first focus on the red scenario. The proposed length for the pair of nodes $2, 3$ is longer than the proposed length for $6, 4$. In this scenario, the resulting labelled history is different to the original one

[1]Starting at the leaves of a rooted tree and continuing towards the root, each pair of leaf nodes coalesces at one interior node. The distance from a pair of leaf nodes to the interior node at which they coalesce is referred to as their *coalescence length* (see Figure 4-1).

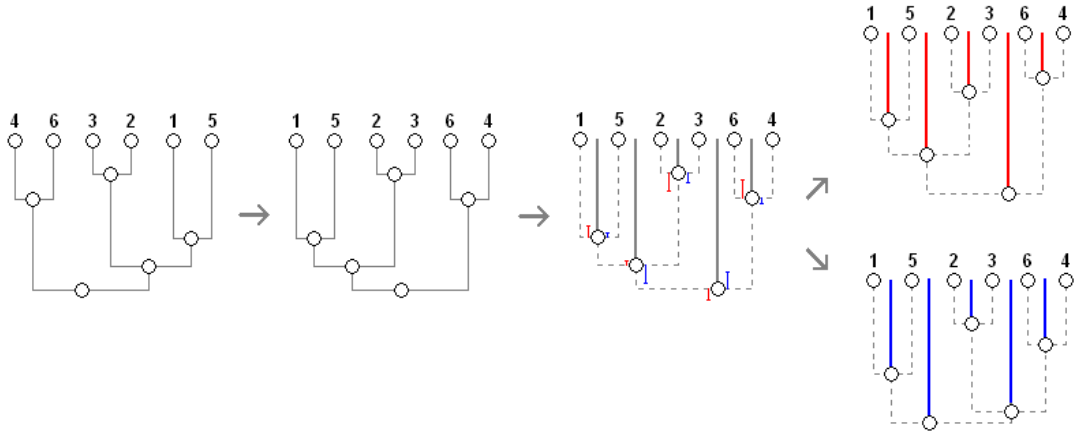


Figure 4-1: In Mau and Newton's [75] proposal, the generation of new coalescence lengths gives rise to two possible scenarios. In red, the candidate lengths are such that a new labelled history is proposed while the tree topology remains unchanged. In blue, the proposed coalescence lengths produce a new tree topology as a side effect.

because the order in which the interior nodes are placed from leaves to root is altered. In the original labelled history, the coalescence of nodes $2, 3$ occurs before that of nodes $6, 4$, but in the proposed labelled history coalescences occur the other way round. In this case, the topology of the tree is conserved and only the order of the interior nodes is modified.

The blue scenario, in contrast, produces new coalescence lengths that have, as a side effect, the generation of a different tree topology. Look at the coalescence length for the pair $5, 2$ on the third tree from left to right (which is the same as for the pairs $1, 2$; $1, 3$ and $5, 3$), and that for the pair $3, 6$ (which extends to the root of the tree). The proposed length under this scenario for the pair $5, 2$ is longer than the proposed length for $3, 6$. This causes a rearrangement in the branching structure and thus, a new tree topology.

The proposal by Mau and Newton makes global rearrangements of coalescence lengths that may or may not change the topology of the tree. A mechanism of this type where both branch lengths and tree are block-updated is hard to control. Branch lengths and trees are distinct parameters whose proposals can be more easily tuned when updated separately. This proposal was eventually called the 'GLOBAL with a molecular clock' by Larget and Simon [63], and is not very commonly used perhaps due to its clock-like restrictions and broad rearrangements.

4.3.2 Larget and Simon's 'LOCAL' proposal

The 'LOCAL' mechanism [63] (also known as 'LOCAL non-clock') starts by randomly selecting one of the $S-3$ internal branches in an unrooted tree. It then labels the two nodes at the end of this selected branch as u and v. This is illustrated in Figure 4-2. The two other neighbouring nodes of u are randomly labelled a and b, while the neighbours of v

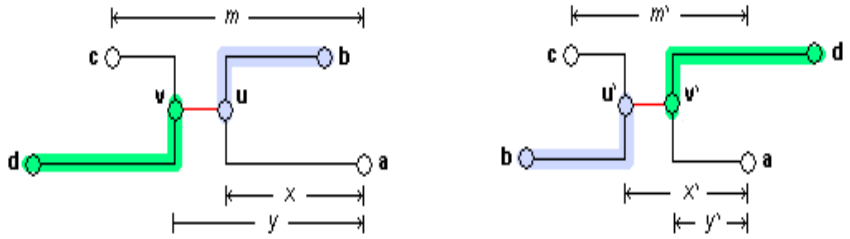


Figure 4-2: The 'LOCAL non-clock' proposal modifies the length of a randomly selected internal branch (shown in red) plus the lengths of the branches leading to nodes a and c; this total length denoted by m. It then detaches either node u or v with equal probability and reattaches it (together with its unchanged subtree b or d) to a point chosen uniformly at random on the path from a to c. If $x' > y'$ (like in this example), the proposal generates a new tree topology, whereas if $x' < y'$ the tree topology remains unchanged. In either case, some branch lengths are also modified.

are randomly called c and d. Suppose that m denotes the distance between nodes a and c. This proposal modifies m by letting $m' = me^{\lambda(r_1 - \frac{1}{2})}$, where $r_1 \sim Uniform(0,1)$ and $\lambda > 0$ is a tuning parameter. The distance between nodes a and u is denoted by x while the distance between a and v by y. The algorithm chooses, with equal probability, either node u or v. If u is selected, it sets $x' = r_2\, m'$ and $y' = ym'm^{-1}$; otherwise $y' = r_2\, m'$ and $x' = xm'm^{-1}$, where $r_2 \sim Uniform(0,1)$.

By changing x and y, this algorithm may also modify the topology of the tree. If $x' > y'$ for instance, the tree topology changes as u' becomes a neighbour of c, and v' becomes a neighbour of a. If $x' < y'$, the tree topology remains unchanged. In either case only the branch lengths in the locality of u and v are modified, hence the proposal's name. (Notice how, in Figure 4-2, the subtrees extending from u' to b, and from v' to d do not suffer any alteration.)

This is a popular proposal that has been implemented in several phylogenetic packages, including BAMBE [108], MrBayes [51, 102] and PhyloBayes [64]. However, there are situations in which the block-update of tree and branch lengths may cause slow mixing of the chain. When inferring the phylogenetic history of a group of taxa, it is not uncommon to find cases where most of the posterior mass is allocated to only a few trees in the tree space while all other trees have negligible posterior support. This typically occurs when the observed DNA sequences belong to well-distinguished species (e.g. [65, 112]). In Chapter 7, for example, we will present the analysis of an alignment containing the DNA sequences of nine primates. In that analysis, a single tree is found to have posterior mass of 0.95, even when the tree space contains $135\,135$ trees in total. In a case like this candidate trees are frequently rejected during simulation due to most of these trees being unsupported by the data. When an MCMC sampler updates both tree and branch lengths *en bloc*, candidate branch lengths may be unfairly rejected in most iterations as a result of unsupported trees being proposed. In contrast, if only branch lengths are updated at a given step (without

modifying the tree), the rate at which candidate branch lengths are accepted should increase (subject to adequate tuning of the proposal) while a proposed tree, generated at a separate step, will be legitimately rejected at most iterations.

Moreover, when estimating more complex models, it is essential to break down the mechanism into simpler blocks to make the path for the chain easier to control. This justifies the need for a new strategy that updates tree and branch lengths at different steps. Nevertheless, the LOCAL proposal could be useful when interested in generating large steps, for example, during the burn-in period of a chain.

4.4 Moves for the homogeneous phylogenetic model

The MCMC sampler implemented in this thesis involves a number of different steps:

(a) updating the phylogenetic tree ϕ;
(b) updating a branch length t_h, for $h = 1, \ldots, (2S-3)$;
(c) updating the substitution rates $\boldsymbol{r}$;
(d) updating the stationary probabilities $\boldsymbol{\pi}$.

One complete pass over these six moves is referred to as an *iteration* and is the basic time-step of our algorithm. Therefore, a single iteration consists of a step to update the phylogenetic tree, $(2S-3)$ steps to update all individual branch lengths separately, one step to generate new substitution rates, and one step to propose new stationary probabilities. All these moves are Metropolis-Hastings. We now present the details of each move type together with the corresponding acceptance probability. We also show irreducibility of the move when appropriate.

4.4.1 Updating the phylogenetic tree

A phylogenetic tree ϕ is updated by the *nearest neighbour interchange* (NNI) mechanism. This mechanism was independently introduced in the early 1970s, first by Robinson [98] and later by Moore, Goodman and Barnabas [81]. Let ϕ be a phylogenetic tree on a label set A. The NNI proposal randomly picks one of the $S-3$ interior branches with equal probability and dissolves the four branches that connect to it, together with the selected branch itself (see Figure 4-3(a)). This leaves four isolated subtrees that can be reconnected in three possible ways. The proposal reconnects the subtrees as in Figure 4-3(b) with probability zero (as this leads to the original tree), and as in Figures 4-3(c) or 4-3(d) with probability $\frac{1}{2}$ each.

The proposal distribution of the forward move is calculated as the probability of being at ϕ and proposing a tree ϕ', with the constraint $\phi \neq \phi'$. This is given by $q(\phi, \phi') = \left(\frac{1}{S-3}\right)\left(\frac{1}{2}\right)$. As this proposal is *symmetric* (i.e. the law of evolution from ϕ' to ϕ is the

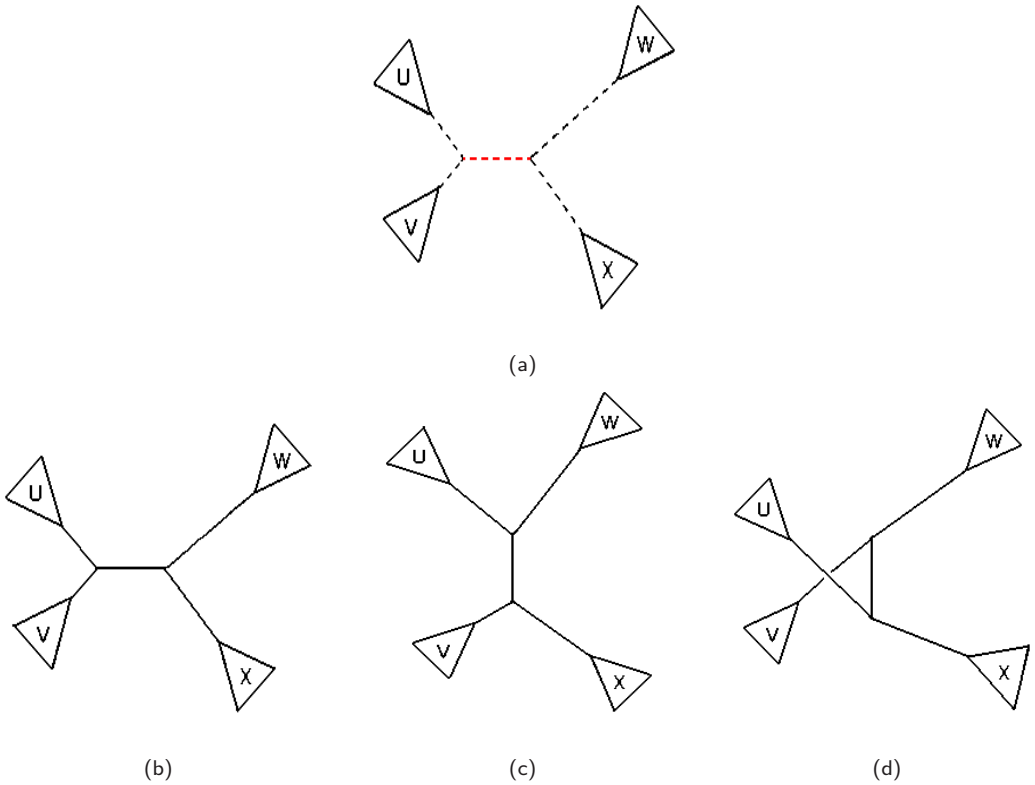


Figure 4-3: (a) In the NNI mechanism for updating a phylogenetic tree, an interior branch is randomly picked and deleted (in red dotted line) together with the four branches that connect to it (in black dotted lines). The four disconnected subtrees that are left (U, V, W and X), are reconnected in one of three possible ways. One way leads back to the original tree (as in (b)), while the other two generate new tree topologies (as in (c) or (d)). The NNI proposal chooses reconnection (b) with probability zero, and either reconnection (c) or (d) with probability $\frac{1}{2}$.

same as from ϕ to ϕ'), the *proposal ratio* $\frac{q(\phi',\phi)}{q(\phi,\phi')}$ simplifies to one (which makes of this proposal a special case of the Metropolis-Hastings algorithm, called a *Metropolis update* [78].)

A candidate tree ϕ' is accepted with probability (from equation (2.9)):

$$\alpha(\phi,\phi') = min\left(1,\ \frac{\cancel{p(\phi')}^{1}}{\cancel{p(\phi)}}\ \frac{L(\phi',\boldsymbol{t},\boldsymbol{\theta}|\boldsymbol{x})}{L(\phi,\boldsymbol{t},\boldsymbol{\theta}|\boldsymbol{x})}\ \frac{\cancel{q(\phi',\phi)}^{1}}{\cancel{q(\phi,\phi')}}\right)$$

$$= min\left(1,\ \frac{L(\phi',\boldsymbol{t},\boldsymbol{\theta}|\boldsymbol{x})}{L(\phi,\boldsymbol{t},\boldsymbol{\theta}|\boldsymbol{x})}\right).$$

Here the prior ratio $\frac{p(\phi')}{p(\phi)} = 1$ as all phylogenetic trees are *a priori* equally likely (Section 3.7), and the acceptance probability simplifies to the minimum of 1 and the likelihood ratio.

4.4.2 Updating a branch length

A new branch length is proposed by one of two mechanisms: the *branch length multiplier* (BLM) or the *branch length normal additive* (BLNA), both of which are well known in the phylogenetics literature. The former is also called the 'proportional shrinking and expanding' move [128] and the latter is alternatively known as the 'sliding window' proposal [128, 51]. Later in this chapter we will show how an MCMC sampler that alternates both proposals, one at each iteration, performs better than a sampler where only one of BLM or BLNA is used.

Branch length multiplier

This move proposes a new length by randomly picking a branch (with all branches equally likely to be picked), whose length is denoted by b. Next, b is multiplied by a quantity m generated from the density

$$f(m) = \frac{1}{\lambda m}, \qquad \frac{1}{\delta} < m < \delta. \tag{4.1}$$

where $\lambda = 2\ln(\delta)$. Simulating from (4.1) can be done by inversion (see for example [82, ch. 5] or [95]). In particular, $m = e^{\lambda(u-\frac{1}{2})}$ generates values from the correct distribution, where $u \sim Uniform(0,1)$. A new branch length is then proposed as

$$\begin{aligned} b' &= b\,m \\ &= b\,e^{\lambda\,(u-\frac{1}{2})} \end{aligned}$$

where $\delta > 1$ is a tuning parameter that influences the step size of the proposal. The density for the forward move (from a current length b to a proposed length b') is equal to the probability density function of b'. Because b' is a transformation of the form $h(m) = b\,m$, its probability density can be calculated as:

$$\begin{aligned} f_1(b') &= f(b'b^{-1})\left|\frac{d}{db'}(b'b^{-1})\right| \\ &= \frac{1}{\lambda b'}, \qquad\qquad \frac{b}{\delta} < b' < \delta\,b. \end{aligned}$$

Similarly, the generating density for the reverse move is calculated as the probability density of b, given by $f_2(b) = \frac{1}{\lambda b}$, where $\frac{b'}{\delta} < b < \delta\,b'$. The proposal ratio thus simplifies to

$$\begin{aligned} \frac{q(b',b)}{q(b,b')} &= \frac{f_2(b)}{f_1(b')} \\ &= \frac{\lambda\,b'}{\lambda\,b} \\ &= m \qquad \text{(since } b' = bm\text{)} \end{aligned}$$

A candidate branch length is accepted with probability

$$\alpha(b, b') = min\left(1,\ \frac{p(b')}{p(b)}\ \frac{L(\phi, \boldsymbol{t}', \boldsymbol{\theta}|\boldsymbol{x})}{L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|\boldsymbol{x})}\ \cancel{\frac{q(b', b)}{q(b, b')}}^{\,m}\right)$$

$$= min\left(1,\ m\, e^{-\beta(b'-b)}\ \frac{L(\phi, \boldsymbol{t}', \boldsymbol{\theta}|\boldsymbol{x})}{L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|\boldsymbol{x})}\right)$$

where $\boldsymbol{t} = (t_1, \ldots, b, \ldots, t_{2S-3})$ is the set of branch lengths. The ratio of the prior distribution $\frac{p(b')}{p(b)} = e^{-\beta(b'-b)}$ is calculated according to the exponential prior for a branch length, with rate parameter $\beta > 0$, specified in Section 3.7.

Branch length normal additive

Let b be the length of a branch that has been randomly selected, with all branches equally likely to be chosen. This move generates a new length by sampling from a normal distribution centred at b and with variance σ^2. This is equivalent to generating from

$$b' = b + \sigma u, \quad \text{with } u \sim N(0, 1) \tag{4.2}$$

where $\sigma > 0$ acts as a tuning parameter that controls the step size of the move and $N(0, 1)$ is the normal distribution with mean 0 and variance 1. Branch lengths are restricted to being greater than zero and any proposed negative value is reflected back at the zero boundary.

The proposal distribution for the forward move is calculated by considering the two ways in which a length b' is obtained; by either sampling a value $b' \geq 0$ at once or by sampling $b' < 0$ and reflecting it back at zero. This is equivalent to generating $b' \geq 0$ from a normal density centred at b and with variance σ^2, or generating $b' < 0$ from a normal density with the same variance but centred at $-b$. Thus, the proposal distribution is given by the sum of two normal densities with the same variance σ^2 and means b and $-b$. Similarly, the reverse move has a proposal distribution that is the sum of two normal densities $N(b', \sigma^2)$ and $N(-b', \sigma^2)$. Thus BLNA is a symmetric mechanism with proposal ratio $\frac{q(b',b)}{q(b,b')} = 1$, and a candidate branch length b' under this proposal is accepted with probability

$$\alpha(b, b') = min\left(1,\ \frac{p(b')}{p(b)}\ \frac{L(\phi, \boldsymbol{t}', \boldsymbol{\theta}|\boldsymbol{x})}{L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|\boldsymbol{x})}\ \cancel{\frac{q(b', b)}{q(b, b')}}^{\,1}\right)$$

$$= min\left(1,\ e^{-\beta(b'-b)}\ \frac{L(\phi, \boldsymbol{t}', \boldsymbol{\theta}|\boldsymbol{x})}{L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|\boldsymbol{x})}\right)$$

where $\boldsymbol{t} = (t_1, \ldots, b, \ldots, t_{2S-3})$ is the set of branch lengths and, as before, the ratio of

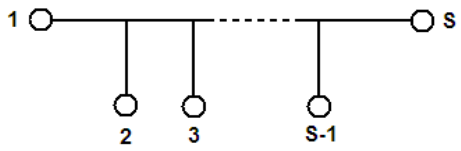


Figure 4-4: The special tree used to show irreducibility of the NNI move. This tree corresponds to a sorted phylogenetic tree on the label set $\{1, \ldots, S\}$, where leaves 1 and S are at either end of one central 'axis' and pendant leaves 2 to $S-1$ are in between. The leaves are sorted in ascending numerical order according to their label.

the prior distribution $\frac{p(b')}{p(b)} = e^{-\beta(b'-b)}$ is calculated according to the exponential prior for a branch length from Section 3.7.

4.4.3 Irreducibility of the tree and branch-length moves

To demonstrate that a Markov chain that moves in the space of phylogenetic trees according to NNI transitions can reach every state from every other state within a finite number of iterations, we show that any tree can be transformed into any other tree in a finite number of NNI moves. The idea is to start at an arbitrary *initial tree* and to perform a series of NNI transformations until the *desired tree* is reached. If we do so via a *special tree*, then we need to show that the initial tree communicates with the special tree and that this, in turn, communicates with the desired tree. Here our special tree is the *sorted tree* shown in Figure 4-4. This tree may be described as one central 'axis', with leaves 1 and S at either end and pendant leaves 2 to $S-1$ in between. The leaves are sorted in ascending numerical order according to their label.

Consider the arbitrary phylogenetic tree on the label set $\{1, \ldots, 5\}$ shown in Figure 4-5(a). In this tree, leaf 4 is not next to leaf 5 but there exists a positive probability of choosing an interior branch such that leaf 4 is moved closer to 5 after an NNI transformation. We achieve this by choosing the branch in red, in Figure 4-5(a), and performing an NNI rearrangement as in Figure 4-5(b). The resulting tree is the one shown in Figure 4-5(c). A second NNI rearrangement, shown in Figure 4-5(d), moves leaf 3 next to leaf 4 and these two steps are enough to reach the sorted tree, which is shown in Figure 4-5(e). Note how, by repeating the same process in the reverse sense, it is possible to reach the arbitrary tree when starting at the sorted tree. It is, therefore, also possible to reach the desired tree from the sorted one.

Now, albeit trivial, we show that any branch length can be reached from any other in a finite number of moves. Let b be the length of a branch selected to be modified by either BLM or BLNA. Under a BLM transformation, let m be the multiplier that scales length b into $b' = bm$. The reverse move, from b' to b, requires the multiplier $m' = 1/m$ so that $b'\,m' = (bm)\left(\frac{1}{m}\right) = b$. Drawing a realisation m from the density (4.1) is straightforward and by basic properties of the real numbers, its multiplicative inverse exists and is unique.

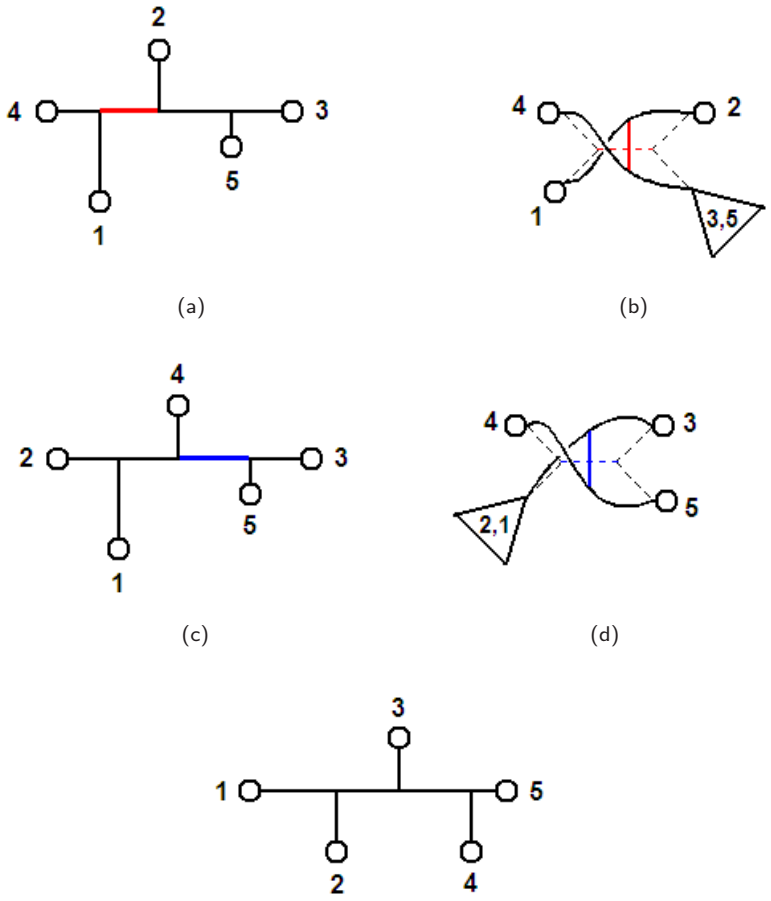


Figure 4-5: Starting at an arbitrary tree on $\{1, 2, \ldots, 5\}$, the chain only requires a finite number of NNI transformations in order to find the sorted tree. (a) By choosing the branch in red and performing an NNI rearrangement as in (b), leaf 4 is placed next to leaf 5, as in (c). After choosing the branch in blue, in (c), and performing a second NNI rearrangement, as in (d), leaf 3 is placed next to leaf 4. (e) The sorted tree is finally reached after only two NNI transformations.

.

Therefore, it is possible to go from b to b' and back again within a finite number of BLM moves.

Similarly, moving from a length b to a length b' according to a BLNA transformation requires drawing either a value b' or $-b'$. Generating either value is possible under a normal distribution. The backward move, from b' to b, requires either a realisation b or $-b$ from a normal distribution, which is also plausible. It is therefore possible to go from b to b' and back again within a finite number of BLNA moves.

4.4.4 Alternating between BLNA and BLM, or using only one?

One of the advantages of an alternating BLNA&BLM mechanism for proposing branch lengths is that one move can be tuned to generate modest steps while the other to produce

replicate	δ	σ
(i)	1.1	0.08
(ii)	1.3	0.06
(iii)	1.5	0.04

Table 4.1: Values for the δ tuning parameter of the BLM move and the σ tuning parameter of the BLNA proposal. These parameters were varied for three different replicates, (i), (ii) and (iii).

bolder ones. In this section, we investigate the performance of single BLNA or BLM updates, compared to a sampler that uses both of them in an alternated manner. To do so, we produced a 6×2500 synthetic DNA alignment with the software package $\mathrm{Seq\text{-}Gen}$ [92] under the GTR model of nucleotide substitution. The values used to generate this alignment are the following: the phylogenetic tree and branch lengths in Newick format[2] are: $(((B : 0.16, D : 0.34) : 0.61, E : 0.2) : 0.53, (C : 0.48, F : 2.14) : 0.35, A : 0.05)$, where $\{A, B, \ldots, F\}$ is the set of leaf-labels and the numbers correspond to the lengths of the branches; the vector of stationary probabilities is $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$ and the substitution rates are $\boldsymbol{r} = (0.140, 0.340, 0.090, 0.008, 0.420, 0.002)$.

In this exercise, we fixed $\boldsymbol{r}$, $\boldsymbol{\pi}$ and the tree to their true values. The target distribution is the joint posterior for branch lengths. We generated candidate branch lengths according to three different methods: (A) from a BLNA proposal; (B) from a BLM proposal; and (C) from an alternating BLNA&BLM scheme. In the alternating BLNA&BLM scheme, candidate branch lengths were generated from the BLNA proposal at even iterations and from the BLM proposal at odd ones. The justification for alternating moves and still converging to the target distribution is given by the fact that if chains P and R have the same stationary distribution, so does PR (see, for example, [44]).

We produced three replicates under each method, varying the tuning parameters of the BLM move (δ parameter) and BLNA (σ parameter). The settings for these replicates are shown in Table 4.1.

Table 4.2 reports the ergodic averages and estimated integrated autocorrelation times for each of the six exterior branch lengths $t_1, \ldots, t_6$, based on $15\,000$ samples after a burn-in period of $5\,000$ iterations. We only report exterior branch lengths since interior branches are not uniquely labelled and so, in subsequent analyses when the sampler also moves in the space of trees, we will be unable to monitor interior branch lengths across different trees. On the top-right of Table 4.2, we have indicated the replicate number and the same order applies for all branches. Notice the better performance of BLNA (small $\hat{\tau}$) relative to

[2]The Newick format is widely used in phylogenetics for representing trees in computer-readable form. It makes use of the correspondence between trees and nested parentheses. This format is further described in [27, ch. 35] and [21].

BLM in estimating $\mathbb{E}(t_1|\boldsymbol{x})$. The results suggest unsuitability of BLM for estimating short branches, which can be further investigated by calculating the expected value and variance of a branch length with respect to the BLM proposal.

In a BLM move, a candidate length is generated as $b' = bm$, where b is the current length and m is a random variable with density function f in (4.1). The expected value and variance of b' are given by

$$\begin{aligned} \mathbb{E}_f(b') &= b\,\mathbb{E}_f(m) \\ &= \frac{b}{\lambda}\left(\delta - \frac{1}{\delta}\right) \end{aligned} \tag{4.3}$$

$$\begin{aligned} Var_f(b') &= b^2\, Var_f(m) \\ &= b^2\left(\frac{1}{2\lambda}\left(\delta^2 - \frac{1}{\delta^2}\right) - \frac{1}{\lambda^2}\left(\delta - \frac{1}{\delta}\right)^2\right), \qquad \frac{b}{\delta} < b' < \delta\, b \end{aligned}$$

where $\lambda = 2\ln(\delta)$ and $\delta > 1$ is a tuning parameter. In the limit $b \to 0$, the expected value and the variance of the candidate length approach $\mathbb{E}_f(b') \to 0$ and $Var_f(b') \to 0$. This produces a phenomenon in which the chain is unable to move away from the zero neighbourhood, which we have dubbed 'zero-stickiness'. A phenomenon like this results in poor estimation performance, since the chain spends several iterations trapped at a small neighbourhood of the state space, producing MCMC samples that are highly correlated to one another.

On the other hand, a candidate branch length is generated from a BLNA proposal as $b' = b + \sigma u$, where $u \sim N(0,1)$ and $\sigma > 0$ is the tuning parameter. Under this proposal, the variance of b' does not depend on the current branch length and the step-size of the move is not influenced but by σ. In estimating $\mathbb{E}(t_6|\boldsymbol{x})$, BLNA performs poorly relative to BLM (see Table 4.2). In other words, when the true branch length is long, BLM outperforms BLNA in all replicates. We believe that this might be due to the fact that the step-size of BLM depends on the current branch length whereas that of BLNA does not. A method that alternates between BLNA and BLM inherits the good properties of both proposals. The results for the combined BLNA&BLM, in Table 4.2, highlight the good estimation performance of such a strategy and justify our preference for alternating between BLNA and BLM when updating branch lengths.

4.4.5 Assessing the tree and branch-length moves

Once having chosen a combined BLNA&BLM strategy as the most convenient approach for updating branch lengths, we now assess the performance of separate tree and branch-length updates relative to a strategy that updates these parameters *en bloc*. In order to

	true length	(A) *BLNA*		(B) *BLM*		(C) *BLNA&BLM*		
		average	$\hat{\tau}$	*average*	$\hat{\tau}$	*average*	$\hat{\tau}$	
t_1	0.05	0.070	**11.01**	0.071	38.01	0.069	**14.06**	(i)
		0.069	**14.18**	0.066	251.84	0.070	**18.30**	(ii)
		0.070	**13.36**	0.070	24.41	0.071	**13.63**	(iii)
t_2	0.16	0.159	10.88	0.157	9.80	0.157	11.76	
		0.158	17.28	0.157	26.73	0.156	19.65	
		0.158	9.69	0.157	9.73	0.157	8.83	
t_3	0.48	0.469	17.40	0.468	22.53	0.469	25.53	
		0.467	21.41	0.465	56.76	0.470	27.70	
		0.470	29.28	0.470	26.95	0.469	30.80	
t_4	0.34	0.310	10.05	0.313	10.05	0.312	11.01	
		0.311	16.52	0.313	19.88	0.312	14.39	
		0.311	11.53	0.312	11.28	0.312	10.52	
t_5	0.20	0.188	9.50	0.188	11.61	0.187	11.08	
		0.188	10.65	0.188	35.64	0.188	17.04	
		0.187	10.53	0.189	9.87	0.187	12.20	
t_6	2.14	2.074	36.17	2.076	**17.21**	2.069	**21.56**	
		2.081	25.97	2.071	**18.12**	2.072	**20.39**	
		2.066	76.32	2.074	**19.89**	2.075	**21.73**	

Table 4.2: The ergodic averages for exterior branch lengths and the estimated integrated autocorrelation time ($\hat{\tau}$), for three different branch-length updating methods: (A) BLNA proposal; (B) BLM proposal; and (C) an alternating BLNA&BLM scheme. For each method, three replicates were performed, each replicate with different tuning parameters. The replicate number is indicated on the top right-hand-side of the table, and this same order applies for all branches. The results correspond to $15\,000$ samples after burn-in. All runs were initialised at the same starting point. The average execution time (across replicates) for the three methods were: (A) $3\,300$, (B) $3\,320$, and (C) $3\,000$; all measured in seconds.

do so, we used the synthetic alignment from Section 4.4.4. We fixed the parameters of the Q-matrix to their true values and made inference on tree topology and branch lengths. The distribution of interest is the joint posterior for ϕ and $\boldsymbol{t}$, whose unnormalised form is

$$\begin{aligned} p(\phi, \boldsymbol{t}|\boldsymbol{x}) &\propto p(\phi)p(\boldsymbol{t})L(\phi, \boldsymbol{t}|\boldsymbol{x}) \\ &\propto e^{-\beta \sum_{h=1}^{9} t_h} \prod_{n=1}^{N} L(\phi, \boldsymbol{t}|x_n) \end{aligned} \tag{4.4}$$

where the exponential term follows from the exponential prior for a branch length, with hyperparameter fixed to $\beta = 10$ (Section 3.7). An unrooted phylogenetic tree relating six organisms has nine branches and we monitored the total branch length, denoted by $T = \sum_{h=1}^{9} t_h$. The reason is because interior branch lengths are not uniquely labelled and it is impossible to monitor them across different topologies throughout MCMC simulation. In our example, the true total length is $T = 4.86$.

In this exercise, we used the phylogenetic software package $\mathrm{MrBayes}$[3] [51, 102], which performs both tempered (Section 2.4.4) and ordinary MCMC runs. We also implemented our own MCMC sampler as a C program called Arbol. The interest here is in the performance of separate and *en bloc* mechanisms, and not in the performance of the programs. ($\mathrm{MrBayes}$ was written by a team of computer scientists and Arbol was not; it would be unfair to make comparisons between them.) The analyses we now present were all run on the same machine and started at the same initial values. In all cases, we generated $20\,000$ samples and discarded the initial $5\,000$ as burn-in.

(1) MrBayes: A standard MCMC run with LOCAL proposal. By default, $\mathrm{MrBayes}$ uses tempered MCMC to improve the mixing of the chain [103]. In this run, however, we specified settings so that no additional chains were used and samples were produced by ordinary MCMC. Branch lengths and tree were updated *en bloc* with a LOCAL proposal. The tuning parameter of LOCAL was set to $\lambda = 0.191$ (this is the default value in $\mathrm{MrBayes}$).

Figure 4-6(a) shows the traceplot of the total branch length for $15\,000$ samples after burn-in. Figure 4-6(b) displays the autocorrelation function of these samples. The ergodic average of the total branch length and the estimated integrated autocorrelation time are reported in Table 4.3. The estimation performance of LOCAL is poor with respect

[3]Incidentally, MrBayes does not allow the user to fix rates of substitution smaller than 0.01. We therefore fixed the rates to $\boldsymbol{r} = (0.14, 0.34, 0.09, 0.01, 0.41, 0.01)$ for all analyses in this section, as opposed to the true $\boldsymbol{r} = (0.14, 0.34, 0.09, 0.008, 0.42, 0.002)$. We postulate that the conclusions derived from the results that we now present hold regardless of the discrepancy in values.

MCMC	*proposal*	$\bar{T}$	$\hat{\tau}(T)$	*time in secs*
(1) standard	LOCAL	4.50	115.54	110
(2) tempered	LOCAL	4.49	69.21	201
(3) standard	NNI / BLNA&BLM	4.50	12.01	1740

Table 4.3: The ergodic average for the total branch length ($\bar{T}$), the estimated integrated autocorrelation time ($\hat{\tau}$) and the computational cost (measured as execution time in seconds) for three different methods: (1) standard MCMC with LOCAL proposal; (2) tempered MCMC with LOCAL proposal; and (3) standard MCMC with NNI / BLNA&BLM updates. The top two rows correspond to runs executed in MrBayes and therefore, their computational costs may be fairly compared. The bottom row reports figures from an analysis in Arbol and its computational cost cannot be fairly compared with analyses (1) and (2).

to independent sampling; the large $\hat{\tau}$ suggests dependence between samples over several iterations.

(2) MrBayes: A tempered MCMC run with LOCAL proposal. In this analysis, branch lengths and tree were updated *en bloc* with a LOCAL proposal, using tempered MCMC with two chains. The tuning parameter of LOCAL was set to the same value as in analysis (1). Figure 4-6(c) shows the trace of T for $15\,000$ samples after burn-in, and Figure 4-6(d) displays their autocorrelation function. The estimation performance of this method, as measured by the estimated integrated autocorrelation time (Table 4.3) has improved relative to method (1). Nevertheless, Table 4.3 also shows that method (2) took nearly twice as long to run as method (1). Using tempered MCMC would only be justified if there was no other alternative that could achieve the same or even greater efficiency at reduced computational cost. There exist such an alternative and we now examine it.

(3) Arbol: A standard MCMC run with NNI / BLNA&BLM proposals. Instead of updating tree and branch lengths *en bloc*, we demonstrate that it is more convenient to update these parameters separately. In this analysis, we used the following moves: an even iteration comprised 9 steps to generate candidates for all branch lengths from the BLNA proposal and one step to generate a candidate tree according to NNI. An odd iteration only differed from an even iteration in that it generated all candidate branch lengths from the BLM proposal instead. The tuning parameters were set to $\delta = 1.5$ for BLM and $\sigma = 0.04$ for BLNA.

Figures 4-6(e) and (f) show the traceplot of T and the autocorrelation function, respectively, for $15\,000$ samples after burn-in. The autocorrelation function highlights the good mixing of the chain. In fact, the estimated integrated autocorrelation time is only 12.01 (see Table 4.3), which makes this method the one with the best estimation performance, relative to methods (1) and (2).

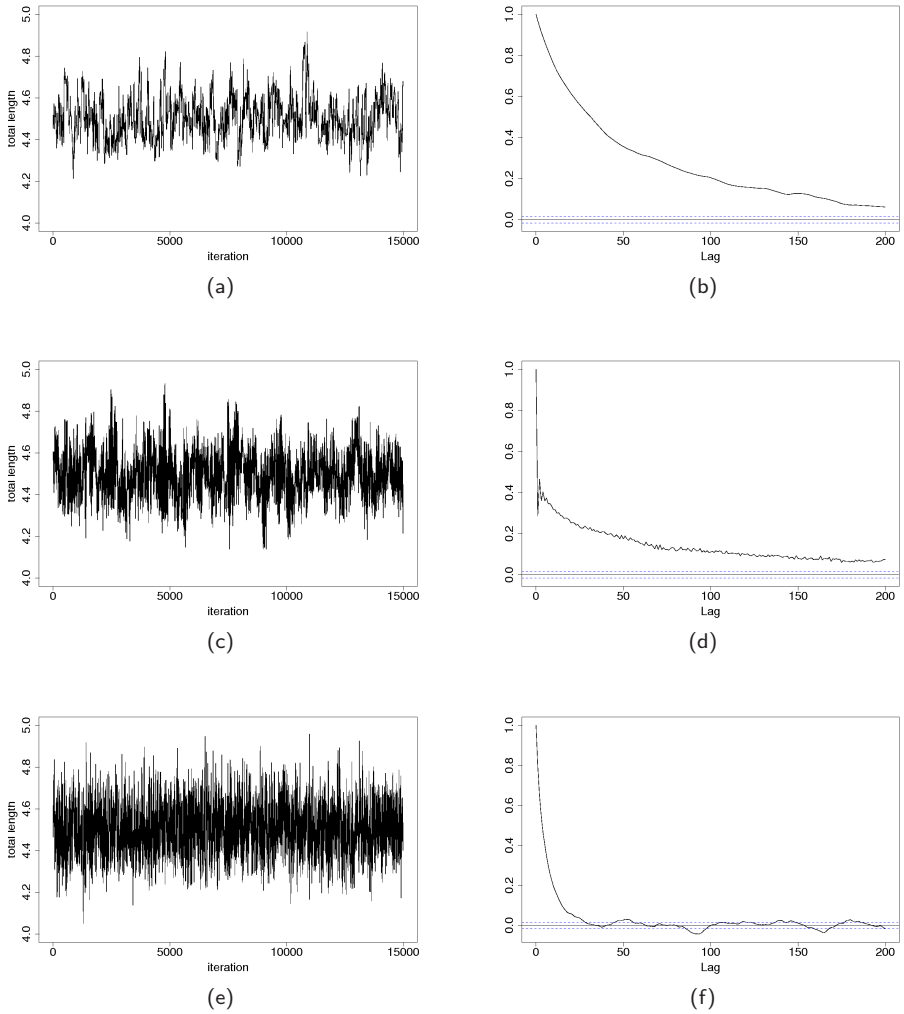


Figure 4-6: Traceplots of the total branch length T, for $15\,000$ samples after burn-in. (a) A standard MCMC with LOCAL proposal; (c) a tempered MCMC with LOCAL proposal ((a) and (c) executed in MrBayes); and (e) a standard MCMC with NNI / BLNA&BLM updates (this latter executed in Arbol). (b),(d),(f) Autocorrelation functions for samples of T obtained with the methods in (a), (c) and (e), respectively.

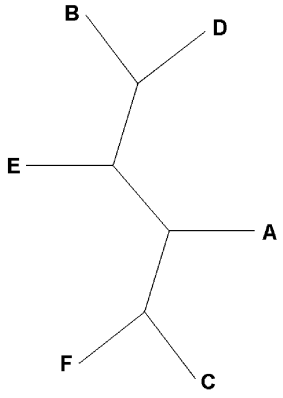


Figure 4-7: Most frequently occurring tree in the MCMC output of analyses (1), (2) and (3), with relative frequency of 100%.

The purpose of this exercise is not to compare the computational costs between programs. However, the fact that method (3) did not require tempered MCMC to achieve the best efficiency indicates that LOCAL is a poor proposal and that tempered MCMC does not make up for it. This is not the first time that the inefficiency of LOCAL is reported. Lakner et al. [62] recently published the results of the performance of seven different proposal mechanisms. They found that proposals producing tree rearrangements as a side-effect of branch length changes (like LOCAL) perform worse than those involving separate tree and branch-length updates.

Figure 4-7 shows the most frequently occurring tree in the MCMC output from analyses (1), (2) and (3). In all cases, this tree topology occurred with relative frequency of 100% and it coincides with the true tree. (The lengths of the branches in this figure are meaningless; the only feature of relevance is the branching structure.) A scenario in which the data support only one tree or very few trees is where a separate update of tree and branch lengths is more advantageous. Moreover, separate updates of tree and branch lengths may also be advantageous when one wishes to make the path of the chain more controllable.

4.4.6 Updating the Q-matrix parameters

Once having presented the details of our tree and branch-length proposals, we now introduce the strategy for updating the parameters of the Q-matrix.

A defective proposal

The GTR model is parametrised by the rates $\boldsymbol{r} = (r_{AC}, \ldots, r_{GT})$ and the stationary probabilities $\boldsymbol{\pi} = (\pi_A, \ldots, \pi_T)$, with both $\boldsymbol{r}$ and $\boldsymbol{\pi}$ constrained to sum to one (Section 3.3.5). Sampling from a Dirichlet distribution is a convenient way of generating a vector in this situation, and this was the approach that Larget and Simon [63] took for generating candidate $\boldsymbol{r}$s and $\boldsymbol{\pi}$s.

Let $\boldsymbol{u} = (u_1, \ldots, u_g)$ be a vector comprising g quantities such that $u_1, \ldots, u_g \geq 0$ and $\sum_{k=1}^{g} u_k = 1$. The *Dirichlet proposal* (DP) generates a candidate $\boldsymbol{u}' = (u'_1, \ldots, u'_g)$ by sampling from a Dirichlet distribution centred at $\boldsymbol{u}$. That is,

$$\boldsymbol{u}' \sim Dir_g(\alpha\, u_1, \ldots, \alpha\, u_g) \tag{4.5}$$

where $\alpha > 0$ is a tuning parameter. A candidate vector $\boldsymbol{u}'$ is accepted with probability

$$\alpha(\boldsymbol{u}, \boldsymbol{u}') = min\left(1,\ \frac{p(\boldsymbol{u}')}{p(\boldsymbol{u})}^{1}\ \frac{L(\phi, \boldsymbol{t}, \boldsymbol{\theta}'|\boldsymbol{x})}{L(\phi, \boldsymbol{t}, \boldsymbol{\theta}|\boldsymbol{x})}\ \frac{q(\boldsymbol{u}', \boldsymbol{u})}{q(\boldsymbol{u}, \boldsymbol{u}')}\right). \tag{4.6}$$

Since $\boldsymbol{u}$ corresponds to either the vector of rates $\boldsymbol{r}$ or stationary probabilities $\boldsymbol{\pi}$, parameter $\boldsymbol{\theta}$ in (4.6) is either $\boldsymbol{\theta} = (\boldsymbol{u}, \boldsymbol{\pi})$ or $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{u})$, depending on whether rates or stationary probabilities are being updated. In acceptance probability (4.6), the prior ratio cancels out given the choice of priors for rates and stationary probabilities in (3.22). The proposal ratio $\frac{q(\boldsymbol{u}', \boldsymbol{u})}{q(\boldsymbol{u}, \boldsymbol{u}')}$ is calculated as the quotient of two Dirichlet density functions.

A drawback of this proposal is that as $u_k \to 0$ the chain may fall into a trap near the zero-boundary as $\mathbb{E}(u'_k) \to 0$ and $Var(u'_k) \to 0$. This can be verified from the functional forms of the expectation and variance of u'_k with respect to the proposal, given by

$$\begin{aligned} \mathbb{E}(u'_k) &= \frac{\alpha\, u_k}{u_0} \\ &= u_k, \qquad\qquad (k = 1, \ldots, g) \\ Var(u'_k) &= \frac{\alpha\, u_k(u_0 - \alpha\, u_k)}{u_0^2(u_0 + 1)} \\ &= \frac{u_k(1 - u_k)}{\alpha + 1}, \qquad (k = 1, \ldots, g) \end{aligned} \tag{4.7}$$

where $u_0 = \sum_{j=1}^{g} \alpha\, u_j = \alpha$. Because the variance of u'_k approaches zero when $u_k \to 0$, the proposal will generate very small steps, all within the neighbourhood of 0. This will result in many instances in the chain path in which the chain is not able to leave the zero-boundary for several iterations.

To illustrate this, we analysed the same synthetic alignment as in Section 4.4.4. The purpose here is to assess the performance of the DP under tempered and ordinary MCMC shemes, as a means of generating candidate $\boldsymbol{r}$s and $\boldsymbol{\pi}$s. The target distribution for this exercise is the joint posterior for rates and stationary probabilities (the tree and branch lengths are fixed to their true values), whose unnormalised form is given by

$$
\begin{aligned}
p(\boldsymbol{r}, \boldsymbol{\pi}|\boldsymbol{x}) &\propto p(\boldsymbol{r})p(\boldsymbol{\pi})L(\boldsymbol{r}, \boldsymbol{\pi}|\boldsymbol{x}) \\
&\propto \prod_{n=1}^{N} L(\boldsymbol{r}, \boldsymbol{\pi}|x_n)
\end{aligned}
\tag{4.8}
$$

where the second line follows from the $Dir(1, \ldots, 1)$ priors for $\boldsymbol{r}$ and $\boldsymbol{\pi}$ given in (3.22). In all cases, the α tuning parameter of the DP proposal was set to 500 and 300 for substitution rates and stationary probabilities, respectively. In order to assess the DP, we used the phylogenetic software package $\mathrm{MrBayes}$ [51, 102] and our own program, Arbol. To assess the computational cost fairly, we recorded the times that the computer needed to complete the runs under the different MCMC methods, while using the same program. The same computer was used for all the runs.

(4) MrBayes: A standard MCMC run with Dirichlet proposal. Vectors $\boldsymbol{r}$ and $\boldsymbol{\pi}$ were updated using the Dirichlet proposal in (4.5). $\mathrm{MrBayes}$ obeys a time-step whereby only one parameter is updated per iteration, meaning that $\boldsymbol{r}$ is updated at odd iterations and $\boldsymbol{\pi}$ at even iterations. Because in our MCMC sampler all parameters are updated at every iteration, we ran a $\mathrm{MrBayes}$ chain with $40\,000$ iterations in total but kept only every other iteration (this standardised $\mathrm{MrBayes}$ output with Arbol output). We then discarded the initial $5\,000$ iterations, leaving a total of $15\,000$ iterations after burn-in.

The sampled values for the substitution rate r_{GT} are shown in Figure 4-8(a). Notice how the chain was not able to leave the zero boundary for several stretches in a couple of occasions. The traceplot suggests slow mixing, and a slowly-mixing chain is not useful for inference [36]. To find out more about the mixing behaviour, it is possible to look at the lag dependence of the chain, as measured by the autocorrelation function, which is shown in Figure 4-8(b). The dependence of the samples is decaying slowly as a function of the lag, suggesting a significant dependence over hundreds of iterations. This confirms that the chain is mixing far too slowly.

How good are the samples produced by this method in estimating the desired expectation $\mathbb{E}(r_{GT}|\boldsymbol{x})$? In order to answer this, we need to estimate the integrated autocorrelation time (Section 2.4.3). This estimate is shown in the first row of Table 4.4. The method has a large $\hat{\tau}(r_{GT})$, which reveals poor performance with respect to independent sampling. Table 4.4 also shows the ergodic average $\bar{r}_{GT}$ and the time that it took to the computer to complete the analysis.

Here we show the sample output from only one run, but similar behaviour was found in several analogous runs. This indicates that the chain is prone to fall into a trap near the zero-boundary when using standard MCMC and updating the parameters of the Q-matrix with a DP proposal.

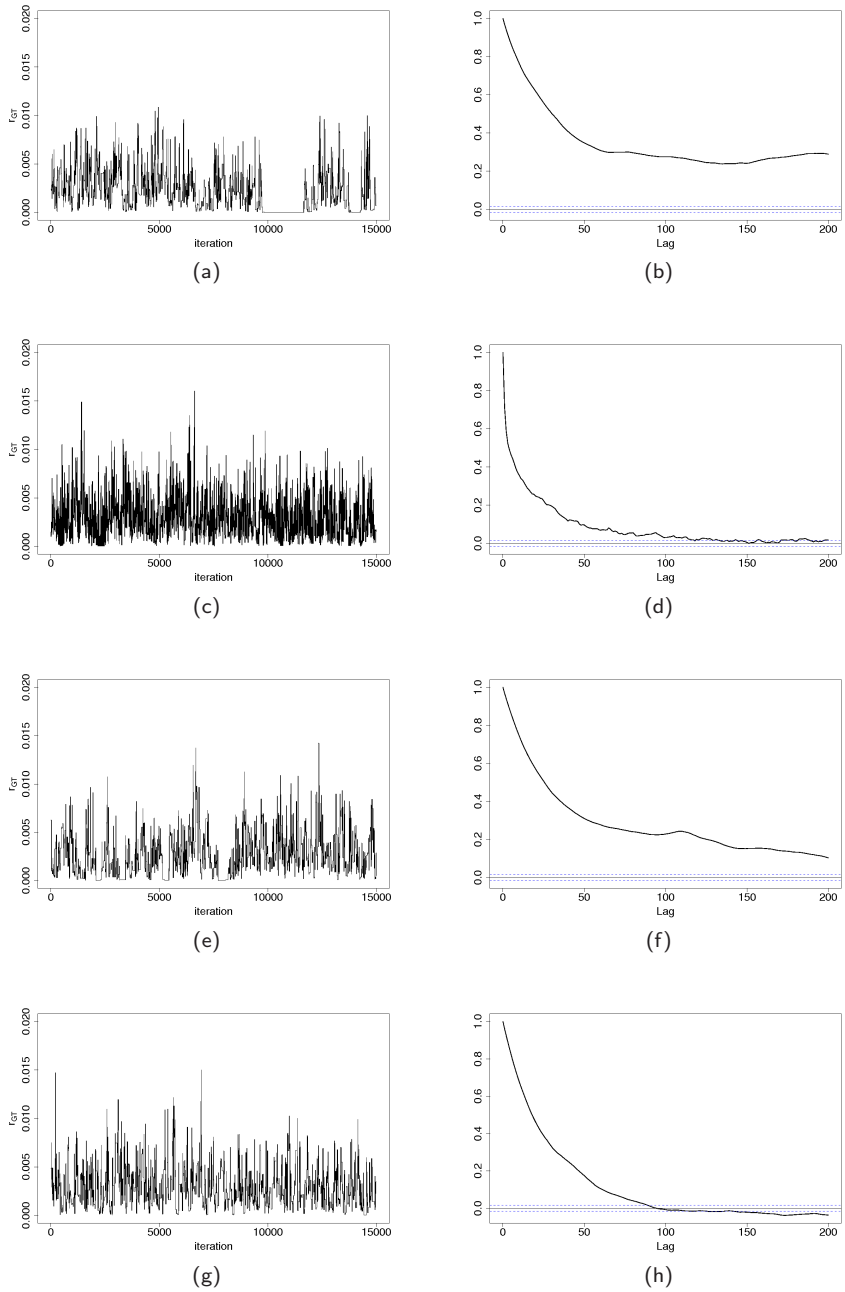


Figure 4-8: Traceplots of r_{GT}, for $15\,000$ samples after burn-in, obtained under (a) a standard MCMC with DP proposal; (c) a tempered MCMC with DP proposal ((a) and (c) executed in $\mathrm{MrBayes}$); (e) a standard MCMC with DP proposal; and (g) a standard MCMC with εDP proposal ((e) and (g) executed in Arbol). (b),(d),(f),(h) Autocorrelation functions for samples of r_{GT} obtained with the methods in (a), (c), (e) and (g), respectively.

(5) MrBayes: A tempered MCMC run with Dirichlet proposal. In this run we used tempered MCMC with one additional 'heated' chain. Candidates for $\boldsymbol{r}$ and $\boldsymbol{\pi}$ were generated using DP with the same tuning parameters as in run (4). We simulated the same number of samples under, otherwise, exactly the same settings and starting values as above. The trace of sampled r_{GT} is shown in Figure 4-8(c), and the autocorrelation function in Figure 4-8(d). The 'zero-stickiness' has disappeared from the chain path and its mixing behaviour seems to have improved. The estimated integrated autocorrelation time is reported in the second row of Table 4.4. The estimation performance of this method is greater than the standard run but the time required to complete the analysis has nearly doubled because of the addition of an extra chain.

The use of the sophisticated and computationally expensive tempered MCMC would be justified only if there was no other method that could achieve similar efficiency with lower computational cost; but there is. In the following section, an improved proposal for updating the parameters of the Q-matrix is presented. Before doing so, though, it is necessary to first record the computational cost of running a chain in Arbol so that we can use this value to compare the cost of other runs.

(6) Arbol: A standard MCMC run with Dirichlet proposal. In this run we used ordinary MCMC to simulate from the posterior. Candidate values for $\boldsymbol{r}$ and $\boldsymbol{\pi}$ were generated from the Dirichlet proposal in (4.5). The length of the chain was set to $20\,000$ iterations and the initial $5\,000$ samples were discarded as burn-in. We used the same settings as for the previous runs and started the chain at the same values as before. The traceplot of rate r_{GT} is shown in Figure 4-8(e). As in analysis (4), the plot exhibits a number of instances in which the chain is trapped near zero for several iterations, due to the 'zero stickiness' caused by the Dirichlet proposal. The slow decay of the autocorrelation function, in Figure 4-8(f), highlights the appreciable dependence between samples over hundreds of iterations and, therefore, the slow mixing of the chain.

This analysis does not reveal anything new about the MCMC method as it is equivalent to run (4) above. Nevertheless, it is necessary to record the time that the computer needed to complete this run if we are to make fair comparisons between different analyses performed in Arbol. Table 4.4 reports the ergodic average $\bar{r}_{GT}$, estimated integrated autocorrelation time $\hat{\tau}$ and required running time for this analysis. The smaller $\hat{\tau}$ with respect to run (4) is only due to the fewer iterations in which the chain remained trapped near zero for this particular realisation of the chain (compare Figures 4-8(a) and (e)). The numbers in this table also indicate that a run in Arbol takes around 7 times longer than an equivalent run in $\mathrm{MrBayes}$ (compare running times for analyses (4) and (6) in Table 4.4).

MCMC	*proposal*	$\bar{r}_{GT}$	$\hat{\tau}(r_{GT})$	*time in secs*
(4) standard	Dirichlet	0.0022	474	51
(5) tempered	Dirichlet	0.0029	46	92
(6) standard	Dirichlet	0.0026	137	354
(7) standard	εDirichlet	0.0029	52	366

Table 4.4: The table shows the ergodic average of the rate of substitution r_{GT}, the estimated integrated autocorrelation time ($\hat{\tau}$) and the computational cost (measured as time needed to complete the analysis, in seconds) for three different methods: (4),(6) standard MCMC with DP proposal; (5) tempered MCMC with DP proposal; and (7) standard MCMC with εDP proposal. The top two rows correspond to runs executed in MrBayes which are, therefore, comparable in computational costs. The bottom two rows report figures from analyses performed in Arbol and their computational costs may be compared.

An improved, novel proposal

The previous section examined the DP as a way of generating candidate $\boldsymbol{r}$s and $\boldsymbol{\pi}$s, and found cases in which the chain falls into a trapping state and is unable to leave the zero neighbourhood for several iterations. When this occurs, the mixing of the chain is negatively affected. The software package MrBayes alleviates this problem by using tempered MCMC, which increases the computational cost of the run by the number of extra chains used. We now present a novel proposal that avoids trapping states at no extra computational cost.

Our proposal generates a new vector by sampling from a Dirichlet distribution centred at the current value but with an ε-shift. If the current vector is $\boldsymbol{u} = (u_1, \ldots, u_g)$, a candidate $\boldsymbol{u}'$ is generated from

$$\boldsymbol{u}' \sim Dir_g(\alpha(u_1 + \varepsilon), \ldots, \alpha(u_g + \varepsilon)) \tag{4.9}$$

where $\varepsilon > 0$ is a small quantity that positively shifts the centre of the Dirichlet distribution and, as before, $\alpha > 0$ is a tuning parameter that controls the variance of the proposal. The probability of accepting a candidate point $\boldsymbol{u}'$ is given by (4.6), with the proposal ratio calculated as the quotient of two Dirichlet density functions with parameters according to (4.9). The expected value and variance of u'_k with respect to the new proposal, are

$$\begin{aligned} \mathbb{E}(u'_k) &= \frac{\alpha(u_k + \varepsilon)}{u_0} \\ &= \frac{u_k + \varepsilon}{1 + 6\varepsilon}, \qquad\qquad (k = 1, \ldots, g) \end{aligned} \tag{4.10}$$

$$
\begin{aligned}
Var(u'_k) &= \frac{\alpha\,(u_k+\varepsilon)\big(u_0-\alpha\,(u_k+\varepsilon)\big)}{u_0^2\,(u_0+1)} \\
&= \frac{\alpha(u_k+\varepsilon)\big(\alpha(1+g\varepsilon)-\alpha(u_k+\varepsilon)\big)}{\alpha^2\,(1+g\varepsilon)^2(\alpha(1+g\varepsilon)+1)} \\
&= \frac{(u_k+\varepsilon)(1-u_k+(g-1)\varepsilon)}{(1+g\varepsilon)^2\,(\alpha+g\alpha\varepsilon+1)}, \qquad\qquad (k=1,\ldots,g)
\end{aligned}
\tag{4.11}
$$

where $u_0=\sum_{k=1}^{g}\alpha(u_k+\varepsilon)=\alpha(1+g\varepsilon)$. In the limit $u_k\to 0$, the expectation $\mathbb{E}(u'_k)\to\frac{\varepsilon}{1+g\varepsilon}$ and the variance $Var(u'_k)\to\frac{\varepsilon(1+(g-1)\varepsilon)}{(1+g\varepsilon)^2\,(\alpha+g\alpha\varepsilon+1)}$, which prevents the chain from falling into a trapping state. The introduction of a shifting parameter ε is a simple way of improving the mixing of the chain without resorting to sophisticated tempered schemes that create extra computational burden. An obvious concern, however, is about the performance of the proposal at the limiting situation in which the value of ε approaches the true r_{GT}. In the following section, we investigate such a scenario but before, we test the εDP in a case where ε is much smaller than the true $r_{GT}=0.002$.

(7) Arbol: A standard MCMC run with εDirichlet proposal. Candidate $\boldsymbol{r}$s and $\boldsymbol{\pi}$s are generated from the εDP in (4.9), with the α tuning parameter fixed at the same value as in method (6) and the shifting parameter set to $\varepsilon=1\times10^{-5}$. The chain was started at the same values and run under exactly the same settings as in (6).

Figure 4-8(g) shows the trace of sampled r_{GT} based on $15\,000$ draws after burn-in. Notice that the 'zero-stickiness' that the chain encountered under a standard MCMC is successfully avoided. The plot of the autocorrelation function, in Figure 4-8(h), highlights the rapid mixing of the chain and $\hat{\tau}(r_{GT})$, in Table 4.4, confirms that the estimation performance of this method is greater than that of method (6). In fact, $\hat{\tau}$ is nearly as good as the one achieved in method (5). The extra computational burden of this method is negligible (compare the two bottom rows in Table 4.4) which, together with its relative accuracy for estimating $\mathbb{E}(r_{GT}|\boldsymbol{x})$, makes it the best of the alternatives presented.

4.4.7 Sensitivity of the novel εDirichlet proposal

How sensitive is the ε-corrected proposal to the choice of ε? We will try to answer to this question by performing a number of runs, each with a different value for ε, and comparing the estimated integrated autocorrelation times for the chain of sampled r_{GT}. The target distribution for this exercise is the joint posterior for rates of substitution and stationary probabilities in (4.8). The tree and branch lengths were fixed to their true values. Substitution rates and stationary probabilities were updated with an εDP proposal, with tuning parameters set to the same values as in methods (6) and (7) – except for ε.

ε	$\bar{r}_{GT}$	$\hat{\tau}(r_{GT})$
0	0.0027	$\mathbf{137}$
1×10^{-8}	0.0028	71
1×10^{-7}	0.0029	70
1×10^{-6}	0.0029	80
1×10^{-5}	0.0029	$\mathbf{52}$
1×10^{-4}	0.0029	60
1×10^{-3}	0.0028	68
2×10^{-3}	0.0029	66
3×10^{-3}	0.0027	$\mathbf{156}$

Table 4.5: Ergodic averages for rate r_{GT} and estimated integrated autocorrelation time $\hat{\tau}$ for $15\,000$ samples after a burn-in. These values are reported for nine runs, where substitution rates and stationary probabilities are generated from an εDP, each run with a different ε-parameter.

Table 4.5 displays the ergodic averages of rate r_{GT} and the estimated integrated autocorrelation times corresponding to $15\,000$ samples after $5\,000$ iterations of burn-in. The worst-performing samplers are those with $\varepsilon = 0$ (which corresponds to the defective DP) and $\varepsilon > 0.002$ (which corresponds to a scenario where ε is greater than the true r_{GT}). The bad estimation performance of the former is due to the 'zero-stickiness' found in the path of the chain and to the failure of DP to move the chain out of the zero neighbourhood. The poor performance of the latter is because the mechanism keeps proposing large steps (steps shifted by a far too large ε) that are often rejected. Table 4.5 also shows that the method with $\varepsilon = 1 \times 10^{-5}$ performs the best as it yields the lowest estimated $\hat{\tau}$.

Based on these results, we conclude that an adequate tuning of the εDP requires a number of initial exploratory runs. By performing a few short runs one could find out if certain values of ε cause bad estimation performance. If for some value of ε the estimated $\hat{\tau}$ is as bad as for the uncorrected DP ($\varepsilon = 0$), one can suspect of ε having exceeded the true parameter value and, therefore, being a bad choice.

We have not examined in detail the effect of α on the performance of the sampler but we know, from equation (4.11), that as α increases the variance of the proposal decreases. A proposal with a small α will generate large steps that will frequently be rejected. This will cause instances in the chain path where the chain does not move and, therefore, bad estimation performance. On the other hand, a proposal with large α will move in small steps and will generally have a high acceptance rate. Nevertheless, the chain will also mix slowly. As before, the α parameter of the εDP proposal has to be adequately tuned to avoid these two extremes.

4.5 Inference on trees

The MCMC sampler presented in this chapter generates a sample from the target distribution to estimate the homogeneous phylogenetic model. Consider a sample $\phi^{(1)}, \ldots, \phi^{(M)}$ of trees from an MCMC run of length M after burn-in. (Here $\phi^{(i)}$ represents the tree sampled at iteration i.) This sequence of trees can be used to count the number of times that a particular tree was sampled throughout the run. The frequency count for a tree, divided by the length of the run M, gives the *relative frequency* for that tree. In particular, we choose the tree with the highest relative frequency as the single 'most likely' tree, given the data and the model. We call this 'most likely' tree the *maximum a posteriori*, or MAP, tree.

In order to produce frequency counts of sampled trees, we have coded in C a separate program based on Diaconis and Holmes' [17] unique-tree-labelling algorithm. This algorithm assigns unique labels to the interior nodes of the tree, according to the identity of the descendants of each interior node. The unique labelling has allowed us to identify each tree with a unique integer and thus keep track of visited trees.

We believe that reporting the MAP tree is not ideal as several other competing trees might have approximately equal posterior support. A fairer treatment of the uncertainty in the data would be to report the *consensus tree*. A consensus tree includes the common elements among the stream of sampled trees during the MCMC run. There exist different consensus methods but, in general, they all count the number of times that a subtree is sampled during the run and assemble a tree from the most frequently sampled subtrees. A good reference for consensus methods is [27, ch. 20]. We have not attempted summarising the posterior for trees by using a consensus tree. However, such an extension would be straightforward as a consensus tree would be generated from the MCMC output that our sampler already produces.

4.6 Discussion

This chapter has presented novel MCMC methodology for estimating the parameters of the homogeneous phylogenetic model. Our strategy performs separate updates of tree and branch lengths. We have proved that this improves the mixing of the chain in some applications. For instance, whenever only one, or very few, trees in the tree space are supported by the data, a mechanism that updates tree and branch lengths *en bloc* will frequently reject both proposed trees and branch lengths. Trees will be legitimately rejected whereas branch lengths might not. The frequent, unreasonable rejection of branch lengths will cause slow mixing of the chain.

In this chapter we also demonstrated how the (non-corrected) DP proposal is unable to prevent the chain from falling into trapping states. Some commercially-available phyloge-

netic programs alleviate this problem by using computationally-expensive MCMC methods, such as tempered MCMC. We proposed a simple ε-correction that solves the 'zero-stickiness' at no extra computational cost. Under a standard MCMC method, our εDP proposal achieves more than twice more accuracy than a DP proposal with negligible increase in computational burden (Table 4.4; compare methods (6) and (7)). Similarly, a standard MCMC method with εDP updates is nearly as accurate as a computationally-prohibitive tempered MCMC method with DP updates (Table 4.4; compare methods (5) and (7)).

The tests carried out in this chapter showed that a method that alternates between BLNA and BLM for updating the branch lengths performs better than a method that uses only one of them (Table 4.2). Similarly, the tests on the sensitivity of εDP to the choice of ε suggested that one should run several exploratory chains before fixing ε to a value. The values chosen for ε and, more generally, for all other tuning parameters, should be chosen to achieve the best possible estimation performance.

Chapter 5

A phylogenetic mixture model for site classification

5.1 Introduction

This chapter introduces a novel mixture model as a convenient way to classify phylogenetic data. This mixture model is expressed as a superposition of simpler component distributions, each of which conforms to the *homogeneous phylogenetic model* with specific parameter values. A mixture like this accounts for heterogeneity underlying DNA data by including multiple sets of branch lengths and Q-matrices. Furthermore, this novel phylogenetic model allows for a formulation whereby is possible to identify the specific component that generates each DNA observation, providing a natural framework for classification.

Our method has similarities and differences with previous approaches, and in this chapter we have included a discussion about them. One of the main distinctions, for example, is that our approach does not require any prior knowledge about class membership nor any *a posteriori* processing for data classification.

5.2 The pathway to modelling heterogeneous DNA data

The homogeneous phylogenetic model postulates that all sites in the DNA alignment evolve under the same phylogenetic tree, set of branch lengths and Markov process of character substitution. In the 1960s, scientists were already aware that the evolution of DNA is heterogeneous, causing different sites in a DNA alignment to experience mutations at different rates (e.g. [29, 60]). Sites that encode crucial functional information are highly conserved through evolution and undergo substitutions at low rates, while sites that are less functionally constrained may experience replacements at a higher rate. A representative example of DNA heterogeneity was discussed in Section 3.3.3, where it was said that the first, second and third codon positions of a protein-coding sequence experience substitutions at different rates. In particular, the substitution rates usually observe an order r_3, r_1, r_2, from high to

low, where r_i denotes the rate of substitution at codon position i [30]. A second example of heterogeneity in DNA evolution is found in the genetic material of bacteria. Bacterial DNA is structured in such a way that certain pieces experience substitutions at high rates as a means to quickly adapt to new environments, while changes in DNA regions that contain essential functional information could be disastrous and these regions hardly ever observe mutations.

There are several published models that account for heterogeneity in DNA data. In order to present an overview of some of them, we have grouped them into three categories: overall-rate models, change-point models and finite mixture models. This review does not intend to be exhaustive and the reader is referred to the original sources for further details.

5.2.1 Overall-rate models

If there are sites that change at different rates, it is possible to account for this by building a model that includes an *overall-rate* parameter γ_n for each site n. Of course, we do not know in advance which sites have which rate and one possible approach is to estimate this parameter for each site. This would require as many extra variables as there are sites in the alignment, which is far from convenient. A better alternative is to treat γ_n as a 'nuisance' quantity and to integrate it out over all possible values it can take. The probability for site x_n is then given by

$$p(x_n|\phi,\boldsymbol{t},\boldsymbol{\theta}) = \int_0^\infty p(\gamma_n)\, p(x_n|\phi,\boldsymbol{t},\boldsymbol{\theta},\gamma_n) d\gamma_n, \quad \text{independently for } n = 1,\ldots,N \tag{5.1}$$

where $p(\gamma_n)$ is a probability distribution for γ_n. This approach relies on finding a realistic $p(\gamma_n)$ that appropriately models the distribution of the overall-rate. In 1971, Uzzell and Corbin [117] suggested using the gamma distribution as the most convenient model. Golding [40] further studied the overall-rate and concluded that this quantity follows a complex distribution and a gamma model was only a first step towards a better description of their variability. In a gamma model, $p(\gamma_n)$ is assumed to be the probability density function of a gamma distribution with shape and inverse scale parameters α and β, respectively. In order to reduce the complexity of this model, it is usual to fix $\mathbb{E}_p(\gamma_n) = 1$ so that $\beta = \alpha$. The gamma model has been shown to provide a reasonable fit to many datasets (e.g. [123, 125]), however, other probability distributions such as the beta or the lognormal have also been reported to provide a similarly good fit (see [27, ch. 13]).

A difficulty of the gamma model has to do with the need to integrate the likelihood $p(x_n|\phi,\boldsymbol{t},\boldsymbol{\theta},\gamma_n)$ over all possible values of γ_n. As discussed in Section 3.6, the likelihood function for phylogenetic parameters has a rather non-standard form, making the evaluation of the integral in (5.1) computationally intensive (if not unfeasible) for a large number of

DNA sequences. The evaluation of (5.1) thus requires approximate methods, the most popular being Yang's discrete-gamma approximation [125]. Yang proposed the evaluation of the likelihood only at k different values. According to Yang's *discrete-gamma model*, the overall-rate at site n takes values $\gamma_{1,n}, \gamma_{2,n}, \ldots, \gamma_{k,n}$ with probabilities $\omega_1 = \ldots = \omega_k = \frac{1}{k}$. Integral (5.1) is then approximated by the weighted sum

$$\int_0^\infty p(\gamma_n)\, p(x_n|\phi, \boldsymbol{t}, \boldsymbol{\theta}, \gamma_n) d\gamma_n \approx \sum_{j=1}^{k} \omega_j\, p(x_n|\phi, \boldsymbol{t}, \boldsymbol{\theta}, \gamma_{j,n}). \tag{5.2}$$

To find $\gamma_{1,n}, \gamma_{2,n}, \ldots, \gamma_{k,n}$, the gamma density is divided into k regions, each of equal density. Atom $\gamma_{j,n}$ is chosen as the mean value within the jth region.

One of the limitations of overall-rate models is that the distribution of γ_n must conform to the specified $p(\gamma_n)$. If the fit of the distribution is poor, the model will not be an adequate description of the evolutionary process. Moreover, the rate matrix Q is not allowed to adopt different configurations across sites since the model only includes one set of θ-parameters. We suspect that important information is missed by explaining rate heterogeneity in DNA sequences in this way. When analysing real data with the gamma model, Yang himself found quite different estimates for the stationary probabilities at codon positions one and three. He then wrote: "Our models assume one common Q for all the codon positions and are not adequate in this respect [...] we suggest that our analyses of rate variation along the sequence will not be influenced much by this inaccuracy of the models [126, p. 99]".

5.2.2 Change-point models

Suchard, Weiss, Dorman and Sinsheimer [111], interested in extending previous existing approaches to modelling heterogeneous DNA data, developed a Bayesian change-point model that allows both the phylogenetic tree ϕ and the parameters of the rate matrix $\boldsymbol{\theta}$ to vary along the sequences. Their prime concern was detecting recombination[1] within the genome of the human immunodeficiency virus (HIV). They assumed that the sites in the DNA alignment separate into an unknown number of contiguous non-overlapping segments, each segment with a distinctive $\boldsymbol{\theta}$ and ϕ. The model successfully inferred the locations in the alignment where the segmentations occur as well as other parameters of interest, namely the number of change-points, the segment-specific phylogenetic trees, and the segment-specific parameters of the rate matrices. However, the current implementation of this model depends on the HKY85 parametrisation of the rate matrix (Section 3.3.4) and

[1] *Recombination* is an essential mechanism of living organisms for adapting to changing environments. It exists in a number of different forms, depending on whether it pertains to eukaryotes, bacteria or viruses. Viruses reproduce in such a way that they can create progeny that carry the genetic information of different parental variants. In other words, part of the genetic material from parental variant A can be joined to part of the genetic material of parental variant B to generate a new viral variant, called a *recombinant strain*. The genetic material of this recombinant strain is a 'mosaic' of different evolutionary processes corresponding to the different evolutionary pathways that each of the parental strains followed. A DNA alignment of putative recombinant strains is thus expected to be separated into a number of non-overlapping segments, each obeying a different phylogenetic tree [66].

works for a rather modest number of DNA sequences. (Suchard and co-workers originally studied an alignment of four sequences; in [79], they extended the model and analysed an eight-sequence dataset.) Also, the nature of their method does not allow for two or more non-contiguous segments to be allocated to the same class; after a change-point, there is always a jump forwards to a new evolutionary class. Indeed, cases exist in nature where distant segments belong to the same evolutionary class [73] and an improved model must take this into account.

5.2.3 Finite mixture models

Hidden Markov models

An alternative model, formulated by Felsenstein and Churchill [28], accounts for the variation in substitution rates across sites by a *hidden Markov process* that operates along the alignment and assigns overall-rates to sites from a finite pool of values. The overall-rate at site n is determined by a hidden random process that depends on which overall-rate was chosen at site $n-1$. Once these rates have been allocated, each site evolves independently according to that overall-rate. In contrast to other less realistic approaches, this method allows there to be some correlation between the rates of evolution at consecutive sites. Then, if a site evolves under a particular overall-rate we expect neighbouring sites to evolve under that same rate with a higher probability than distant sites.

Felsenstein and Churchill's strategy, as overall-rate models do, assume a common-to-all-site rate matrix Q. An improved treatment of rate variation must allow for a specific configuration of the Q matrix according to the type of evolution each site follows; i.e. allow for site n to evolve according to rate matrix Q_n and site m according to Q_m ($Q_n \neq Q_m$). A disadvantage of Felsenstein and Churchill's hidden Markov model is that possible biases may be introduced by the removal of sites involving gaps in the alignment or other errors resulting in consecutive observable sites not being direct neighbours in nature. The correlation assumption of this model relies on adjacent sites in the observables actually having some direct effect on one another. Moreover, evolutionary processes operating at different codon positions cannot be adequately explained by this model because the positions among codons are correlated at intervals of three and not at consecutive sites (i.e. site n correlates with site $n-3$ and not with $n-1$).

Pagel and Meade's mixtures

In 2004, Pagel and Meade [88] modelled heterogeneity in DNA data by postulating a mixture of several different evolutionary classes operating in nature to generate the observable DNA data. Their model supposes that data at site n arise from a distribution

$$x_n \sim \sum_{j=1}^{k} \omega_j \, p(x_n|\phi, \boldsymbol{t}, \boldsymbol{\theta}_j), \quad \text{independently for sites } n = 1, \ldots, N \tag{5.3}$$

where k is the number of evolutionary classes; $\omega_1, \ldots, \omega_k$ are the mixture proportions ($\omega_j \geq 0$ and $\sum_{j=1}^{k} \omega_j = 1$); and $p(\cdot|\phi, \boldsymbol{t}, \boldsymbol{\theta}_1), \ldots, p(\cdot|\phi, \boldsymbol{t}, \boldsymbol{\theta}_k)$ are the component densities, each indexed by a phylogenetic tree ϕ, set of branch lengths $\boldsymbol{t}$ and a component-specific $\boldsymbol{\theta}_j$ that specifies a rate matrix Q_j.

One of the interpretations of a mixture model is as a representation of a heterogeneous population consisting of classes $j = 1, \ldots, k$ of sizes proportional to ω_j, from which the observations $x_1, \ldots, x_N$ are drawn. Formulation (5.3) asserts that the heterogeneity between classes is limited to the information contained in the rate matrices since the phylogenetic tree ϕ and the set of branch lengths $\boldsymbol{t}$ are assumed to be common to all k mixture components. The rate matrix may be thought of as a qualitative description of the substitution process; different configurations of Q represent distinct *rules* of substitution. In model (5.3), one rate matrix may conform to a substitution process with a noticeable difference between rates of transition (r_{AG}, r_{CT}) and rates of transversion ($r_{AC}, r_{AT}, r_{CG}, r_{GT}$), while another rate matrix may agree with a process where this difference is milder but the proportion of one nucleotide is remarkably higher than the others (e.g. a large π_A relative to π_C, π_G, π_T). Whatever their configuration, rate matrices are standardised so that the expected total rate of substitution is one (Section 3.4.2). As a result, a mixture of Q-matrices like (5.3) does not account for the variability in the *amount* (or *quantity*) of substitution events, but only for the variability in the *rules* of substitution.

Aware of this, Pagel and Meade [88] incorporated the idea of an overall-rate of substitution to their mixture model and devised the rather intricate formulation

$$p(x_n|\phi, \boldsymbol{t}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) = \int_0^\infty p(\gamma_n) \sum_{j=1}^{k} \omega_j \, p(x_n|\phi, \boldsymbol{t}, \gamma_n, \boldsymbol{\theta}_j) d\gamma_n, \quad \text{independently for } n = 1, \ldots, N$$

for the probability of site x_n. They chose $p(\gamma_n)$ to be the probability density function of a gamma distribution and used Yang's approximation [125] to evaluate the integral. It was at this stage that the research underlying this thesis started in 2005. One of its aims was to find a more natural way of modelling both heterogeneity in the *amount* and in the *rules* of evolution, one that did not involve the overall-rate γ_n and the consequent evaluation of the integral.

A drawback of Pagel and Meade's models is that they do not provide any information whatsoever about the membership of a site to a mixture component. In their formulation, all sites are identically modelled by a mixture of k components, which certainly accounts for the heterogeneity in the data but does not tell us which sites evolve according to which component. In a phylogenetic context, mixture components have a direct biological interpretation; components may correspond to different codon positions in a protein-coding

region or to different genes. Learning about the membership of a site to one or another mixture component can bring valuable insights into the evolutionary processes generating the data.

To determine the probability of an observation being generated by a particular component, Pagel and Meade [88] suggested a break down of the mixture into individual components for separate analysis. That is, they first analysed the data under the mixture model to obtain parameter estimates. Then they calculated the probability of observing site x_n given an estimated $\hat{\boldsymbol{\theta}}_j$ and some other estimates. They did this separately for all components $j = 1, \ldots, k$. The θ-estimate that achieved the highest probability was the one that best explained observation x_n. Their strategy required expensive *a posteriori* processing of the data. An improved approach would model the class allocation of x_n jointly with the model parameters and perform inference on both allocations and parameters at once. This is one of the enhancements added to the novel phylogenetic mixture model that is now presented.

5.3 The $Q + t$ mixture model

We have already mentioned the importance of being able to learn about the membership of sites to evolutionary classes. For instance, when analysing a DNA alignment constructed by the combination of several genes, called a *gene concatenation*, the differently-evolving genes may conform to the distinct classes. Classifying DNA sites to different classes allows us to draw conclusions about the evolutionary compatibility or incompatibility between genes. Sites known to belong to different genes that are classified to the same class are evidence of evolutionary congruence between the genes. In contrast, sites known to belong to different genes which are clustered in different groups are an indication of evolutionary incompatibility between those genes. A finite mixture model, through its missing-data formulation, provides an appropriate means for classifying data.

The basic finite mixture model for site x_n is written as

$$x_n \sim \sum_{j=1}^{k} \omega_j \, p(x_n | \phi, \boldsymbol{t}_j, \boldsymbol{\theta}_j), \quad \text{independently for sites } n = 1, \ldots, N \tag{5.4}$$

where k is the number of mixture components; $\omega_1, \ldots, \omega_k$ are the mixture proportions ($\omega_j \geq 0$ and $\sum_{j=1}^{k} \omega_j = 1$); and $p(\cdot|\phi, \boldsymbol{t}_1, \boldsymbol{\theta}_1), \ldots, p(\cdot|\phi, \boldsymbol{t}_k, \boldsymbol{\theta}_k)$ are the component densities. The component-specific parameters are grouped in a set $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k) = \big((\boldsymbol{t}_1, \boldsymbol{\theta}_1), \ldots, (\boldsymbol{t}_k, \boldsymbol{\theta}_k)\big)$. Within a phylogenetic context, model (5.4) asserts that site x_n is generated from a mixture of k different evolutionary processes operating in nature in proportions $\omega_1, \ldots, \omega_k$. The jth evolutionary process, $p(\cdot|\phi, \boldsymbol{\psi}_j)$, is indexed by a common-to-all-component phylogenetic tree ϕ, and a component-specific set of branch lengths $\boldsymbol{t}_j$ and Q-matrix parameters $\boldsymbol{\theta}_j$.

Branch lengths are the product of rate of substitution and time. The length of a branch, therefore, represents the *expected number of nucleotide substitutions* accumulated between the two nodes it connects (Section 3.4.2). In order to facilitate the discussion, let us think of a set of branch lengths $\boldsymbol{t} = (t_1, \ldots, t_{2S-3})$ in terms of its total length, i.e. $\sum_{h=1}^{2S-3} t_h$. Model (5.4) allows there to be k distinct total lengths, each one corresponding to a unique configuration of the set of individual branch lengths. A mixture component with a large total length conforms to an evolutionary process where, on average, branches are accumulating a high number of substitutions, while a mixture component with a small total length corresponds to an evolutionary process where branches, on average, are short. The former component accounts for rapidly evolving sites that undergo substitutions more frequently and the latter agrees with sites that accumulate a lower number of substitutions. Consequently, postulating a mixture model that includes multiple sets of branch lengths and multiple Q-matrix parameters is a natural way of allowing for both heterogeneity in the *amount* and *rules* of evolution. The different sets of branch lengths model the variability in the expected *number* of nucleotide substitutions across sites and the different sets of Q-matrix parameters explain the heterogeneity in the *rules* of evolution.

Furthermore, we adopt a missing-data reformulation of the mixture model in which observation x_n is *augmented* by a quantity z_n, called the *allocation variable*, taking values in the set $\{1, \ldots, k\}$ (Section 2.6.2). The allocation variable is an (unobserved) integer identifying the underlying generating component of site x_n. Conditional on z_n, observation x_n is independently drawn from the density corresponding to the z_nth mixture component. That is,

$$x_n | z_n \sim p(x_n | \phi, \boldsymbol{t}_{z_n}, \boldsymbol{\theta}_{z_n}), \quad \text{independently for sites } n = 1, \ldots, N. \tag{5.5}$$

Expression (5.5) reflects the fact that, once the allocation for site n is known, this observation is 'demixed' and the site is no longer considered as generated by a mixture of k processes but as generated by process z_n. The use of a missing-data formulation allows the complicated structure of the mixture to be decomposed into simpler structures. This approach has long been used in the statistical literature when one of the objectives of the analysis is to classify a set of N observations into k mutually exclusive classes (e.g. [122, 4]). To the best of our knowledge, this is the first time that such a formulation is used within a Bayesian phylogenetics context and it represents one of the major contributions of this work. Throughout, the model specified by (5.4) and (5.5) is referred to as the $Q + t$ *mixture model*.

5.3.1 Model hierarchy and choice of priors

The joint distribution of all unknown quantities is given by

$$
\begin{aligned}
p(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi}, \boldsymbol{x}) &= p(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi}) L(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi} | \boldsymbol{x}) \\
&= p(\boldsymbol{\omega}) p(\boldsymbol{z} | \boldsymbol{\omega}) p(\phi, \boldsymbol{\psi} | \boldsymbol{\omega}, \boldsymbol{z}) L(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi} | \boldsymbol{x}) \\
&= p(\boldsymbol{\omega}) p(\boldsymbol{z} | \boldsymbol{\omega}) p(\phi) p(\boldsymbol{\psi}) L(\boldsymbol{z}, \phi, \boldsymbol{\psi} | \boldsymbol{x})
\end{aligned} \tag{5.6}
$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k)$ is the vector of mixture proportions; $\boldsymbol{z} = (z_1, \ldots, z_N)$ is the vector of allocation variables for sites $1, \ldots, N$; ϕ is the phylogenetic tree; $\boldsymbol{\psi} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ is the set of component-specific parameters; and $\boldsymbol{x} = (x_1, \ldots, x_N)$ is an observed DNA alignment.

In (5.6), the last line follows by imposing further conditional independences so that $p(\phi, \boldsymbol{\psi} | \boldsymbol{\omega}, \boldsymbol{z}) = p(\phi) p(\boldsymbol{\psi})$ and $L(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi} | \boldsymbol{x}) = L(\boldsymbol{z}, \phi, \boldsymbol{\psi} | \boldsymbol{x})$. The latter obeys the missing-data interpretation of the mixture in which, once knowing the allocations $\boldsymbol{z}$, each observation is drawn from its corresponding component.

The likelihood $L(\boldsymbol{z}, \phi, \boldsymbol{\psi} | \boldsymbol{x})$ is calculated as the product of the sampling distribution (5.5), from site 1 to N, as follows (refer to Section 3.6 for details on the computation of the likelihood function):

$$
L(\boldsymbol{z}, \phi, \boldsymbol{\psi} | \boldsymbol{x}) = \prod_{n=1}^{N} L(\phi, \boldsymbol{t}_{z_n}, \boldsymbol{\theta}_{z_n} | x_n). \tag{5.7}
$$

All that remains to specify is the prior distributions on model parameters. The prior distribution on $\boldsymbol{\omega}$ is taken to be the symmetric Dirichlet distribution with parameter 1,

$$
\boldsymbol{\omega} \sim Dir_k(1, \ldots, 1)
$$

which is a proper (albeit unnormalised) prior and implies no prior knowledge about the relative sizes of the mixture components. Conditional on $\boldsymbol{\omega}$, the allocations $z_1, \ldots, z_N$ are assumed independent and identically distributed with probability

$$
p(z_n = j | \boldsymbol{\omega}) = \omega_j, \quad j = 1, \ldots, k. \tag{5.8}
$$

As in the homogeneous model, all phylogenetic trees with label set $(1, 2, \ldots, S)$ are assumed to be equally likely *a priori* (probability (3.20)). The prior distribution for parameters $\boldsymbol{\psi}$ is specified as

$$
p(\boldsymbol{\psi}) = \prod_{j=1}^{k} p(\boldsymbol{t}_j) p(\boldsymbol{\theta}_j) \tag{5.9}
$$

where it is assumed that prior distributions on sets of branch lengths and rate-matrix

parameters belonging to different mixture components are independent. Vector $\boldsymbol{t}_j$ contains the $2S-3$ individual branch lengths for the jth mixture component so that $\boldsymbol{t}_j = (t_{(1,j)}, \ldots, t_{(2S-3,j)})$. The model specifies independent exponential priors on individual branch lengths, with rate parameter β, as in (3.21). Vector $\boldsymbol{\theta}_j$ contains the set of six substitution rates for the jth component, $\boldsymbol{r}_j = (r_{(AC,j)}, \ldots, r_{(GT,j)})$, and the set of four stationary probabilities $\boldsymbol{\pi}_j = (\pi_{(A,j)}, \ldots, \pi_{(T,j)})$ for component j. The model defines independent prior distributions on $\boldsymbol{r}_j$ and $\boldsymbol{\pi}_j$, as in (3.22).

The joint prior distribution can be finally written as:

$$p(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi}) \propto \prod_{j=1}^{k} \omega_j^{N_j} \, e^{-\beta \sum_{h=1}^{2S-3} t_{(h,j)}} \tag{5.10}$$

where $N_j = \sum_{n=1}^{N} I[z_n = j]$ and $I[\cdot]$ is the indicator function taking value 1 when its argument is true and 0 otherwise. (Notice that $\sum_{j=1}^{k} N_j = N$.) The factor $\prod_{j=1}^{k} \omega_j^{N_j}$ follows from distribution (5.8).

Directed acyclic graph of the $Q+t$ mixture model

The hierarchical structure of the $Q+t$ mixture is displayed in Figure 5-1 as a directed acyclic graph (DAG). The DAG is equivalent to the assumption that allows us to decompose the full joint distribution (5.6) into smaller components [5]. The square nodes represent variables which are known, such as observed data or hyperparameters whose values are fixed, while circles indicate the unknowns. The direction of an edge between two nodes represents the direction of the relationship between the two corresponding variables.

Figure 5-1 represents the belief that allocation to a particular mixture component depends on the number and sizes of the components (k and $\boldsymbol{\omega}$, respectively). Similarly, an observation x depends on the allocation variable z but is conditionally independent on the values of $\boldsymbol{\omega}$ and k, given the value of z. This conditional independence between x and parameters $\boldsymbol{\omega}$ and k is represented in the graph by the absence of an edge between nodes $\boldsymbol{\omega}$ and x, and between k and x. In general, the graph tells us that, given the value of the allocation variable, knowing the mixture proportions and the number of components provides no extra information about an observable x. A similar rationale can be followed to interpret the rest of the graph.

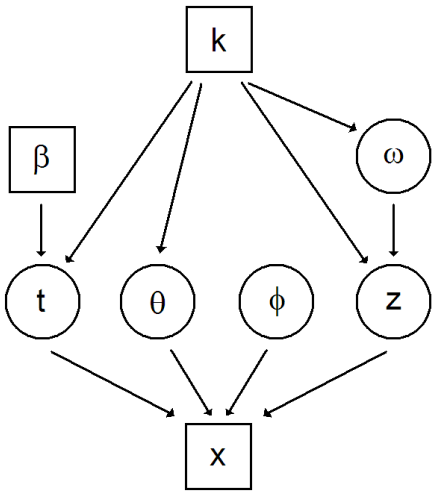


Figure 5-1: Directed acyclic graph of the $Q + t$ mixture displaying the hierarchical dependence structure of the model.

5.4 Number of mixture components

Perhaps one of the most fundamental aspects of a finite mixture model is the number of mixture components k. In order to determine k, one could adopt a fully Bayesian approach and consider k as an unknown parameter in the formulation of the Bayesian model. In this situation, mixtures with a varying number of components would need to be sampled and reversible jump MCMC [41, 42] is commonly used to do so.

An alternative approach, and the one taken in this thesis, is to consider mixtures for a fixed number of components separately and, at a later stage, use some criterion to decide on the most suitable value for k. The standard Bayesian solution decides on the value of k by calculating the *Bayes factor* for two competing models with different k. The Bayes factor B_{21} is a summary of the evidence provided by the data in favour of one model as opposed to another [58]. It is defined as the ratio of the marginal likelihood under model H_2 to the marginal likelihood under a second model H_1, when neither model is favoured over the other *a priori*. Then,

$$B_{21} = \frac{p(\boldsymbol{x}|H_2)}{p(\boldsymbol{x}|H_1)} = \frac{\int p(\vartheta_2|H_2)p(\boldsymbol{x}|\vartheta_2, H_2)d\vartheta_2}{\int p(\vartheta_1|H_1)p(\boldsymbol{x}|\vartheta_1, H_1)d\vartheta_1}$$

where ϑ_i is the vector of parameters of model H_i and $p(\vartheta_i|H_i)$ is its prior density. Once B_{21} has been calculated, it can be interpreted as strong evidence against model H_2 or strong evidence for it (with a range of levels in between), according to a scale provided by Kass and Raftery [58].

The marginal likelihood, $p(\boldsymbol{x}|H_i)$, is the key quantity needed in the computation of the Bayes factor and, for complex problems, it is unusual to be able to calculate it analytically.

Kass and Raftery [58] provide a thorough account of available methods for its estimation, some of which are based on MCMC samples from the posterior distribution for ϑ. Chib and Jeliazkov [15] developed a method in the context of MCMC output produced by the Metropolis-Hastings algorithm. Nevertheless, estimation of $p(\boldsymbol{x}|H_i)$ in complex, high-dimensional scenarios may produce highly inaccurate answers, as reported by Raftery [91].

In [112], Suchard, Weiss and Sinsheimer use Bayes factors to test for a phylogenetic model with multiple Q-matrices against a model with a single Q-matrix. In order to estimate the marginal likelihood they make important simplifying assumptions, such as model H_1 being nested within model H_2. This prevents them from testing models that condition on the tree and branch lengths as parameters of interest. The set of branch lengths does not maintain consistent definition between different topologies – interior branch lengths are not uniquely labelled – and models with different tree topologies are non-nested models.

Given the underlying complexity of Bayes factors in a phylogenetic context, the criterion taken in this thesis was to choose the largest value of k such that all the components are different and all the associated mixture proportions are non-zero. This can be achieved by inspection of the MCMC output. Some prior knowledge about the nature of the observed DNA sequences can also aid in hypothesising a value for k. For instance, if analysing a concatenation of genes, one possible value for k to try is the number of different genes in the alignment. The computational cost of a phylogenetic mixture with large values of k is high, so it is desirable to begin with a small value and assess its adequacy before moving to a larger k.

5.5 Discussion

5.5.1 Some advantages of mixture modelling

This chapter introduced the novel $Q + t$ mixture model as a convenient means to classify DNA observations into k distinct evolutionary classes. If there are k classes generating observable x_n, it is natural to suppose that there is a way to augment x_n with an (unobserved) allocation variable z_n that specifies the mixture component from which the observable is generated. Once z_n is available, the observable is no longer generated by a mixture but by the specific component that the allocation indicates. As a result, the parameters of each component can be simulated conditionally on the allocations, taking into account only the sites that have been allocated to that component at a particular MCMC iteration. In most instances, the calculation of the acceptance probability of a Metropolis-Hastings step greatly simplifies as a direct consequence of this.

Mixtures of distributions are known to provide the means to model quite complex distributions with few parameters and a high level of accuracy. Advocates of the *gamma model*, however, may argue that heterogeneous DNA data can be satisfactorily explained

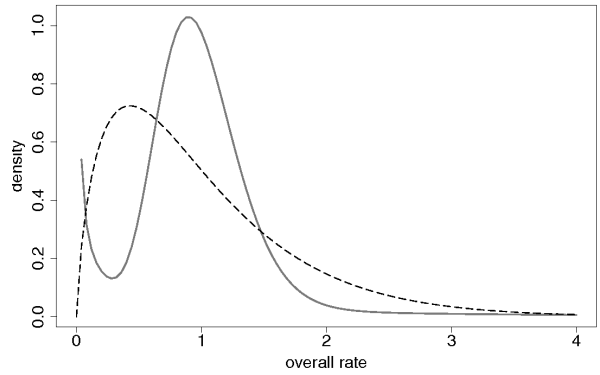


Figure 5-2: Estimated density (dotted curve) when the gamma model is used to analyse an alignment in which the overall-rate variation follows a non-standard bimodal distribution (solid curve). The estimated curve is an inadequate description of the overall-rate heterogeneity.

by building a model that includes an overall-rate for each site, sampled from a gamma distribution. In comparison to the homogeneous formulation, the gamma model accounts for heterogeneity in an elegant and convenient way by adding only one extra parameter (the α-shape parameter of the gamma distribution). But convenience and elegance lose their appeal when trying to deal with more complex scenarios.

Consider a DNA alignment in which the overall rates follow a non-standard bimodal distribution, like the one shown in Figure 5-2 as a solid curve. An analysis with the gamma model fits a standard gamma distribution and produces an estimated density like the dotted curve displayed in the same figure. In situations like this, the gamma approach is not satisfactory as it misses important information about the heterogeneity underlying the data. A mixture model that postulates two classes of overall-rates would provide a more adequate explanation of the process generating these data.

The number of parameters included in a mixture model is a reasonable compromise between realism and complexity whenever the data are not adequately modelled by a cheaper formulation. In cases in which the observations are suspected to follow a simpler (albeit still heterogeneous) generating process, alternative approaches, such as the gamma model, may be preferred. Formal statistical tests for choosing between competing phylogenetic models are discussed by Felsenstein [27, ch. 19] and have been used within a context of phylogenetic mixtures by Pagel and Meade [89]. These tests are based on the Akaike information criterion [1] and the Bayesian information criterion [104]. Suchard et al. [112] discuss model selection via Bayes factors.

5.5.2 Classification criterion

The $Q + t$ mixture model postulates that the source of heterogeneity in phylogenetic data is the lengths of the branches, $\boldsymbol{t}$, and the parameters of the Q-matrix, $\boldsymbol{\theta}$. Both $\boldsymbol{t}$ and

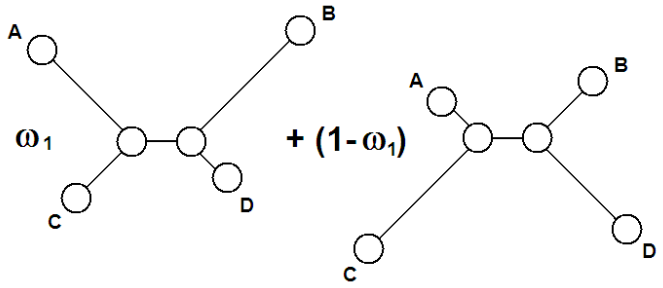


Figure 5-3: A phenomenon, known as *heterotachy* [69], in which a proportion ω_1 of sites evolve under a tree with long branches leading to species A and B, while the remaining sites evolve under a tree with shorter branches leading to these same species.

$\boldsymbol{\theta}$ are related to the rate at which nucleotides experience substitutions; the former in a *quantitative* way while the latter in a *qualitative* manner. Multiple Q-matrices allow for different configurations of the Markov substitution-process across classes while multiple sets of branch lengths allow for varying rates of substitution both across classes and between species. The variation of rates across species is illustrated in Figure 5-3. Here a proportion ω_1 of sites evolve under a tree with a large number of substitution events in the branches leading to species A and B (long branches). The remaining sites evolve under a tree with fewer substitutions along these lineages (short branches leading to A and B). This phenomenon, which causes different sites to have differently 'accelerated' branches, was named by Lopez, Casane and Philippe [69] as *heterotachy*.

The principle of heterotachy states that the rates along branches leading to different species in the tree can vary among sites. Thus, some sites have long branches leading to certain species while these same branches are shorter in other sites. This phenomenon has been repeatedly reported in empirical phylogenetic studies (e.g. [68, 90, 67, 100]) and it is believed to play a part in evolutionary processes. A heterogeneous formulation, such as the gamma model, fails to account for heterotachy since the overall-rate of a site equally affects all the branches across the tree, hence the term 'overall'. A mixture model that includes multiple sets of branch lengths captures this variability and allows for different sets of lengths to adopt different configurations, i.e. allows for heterotachous classes. Indeed since the beginning of this thesis a number of research groups have independently proposed mixtures on sets of branch lengths as a way of effectively modelling heterotachy in phylogenetic data [89, 77, 61]. A brief review of these mixtures is provided below.

The choice of a common phylogenetic tree ϕ across all the components of the $Q+t$ mixture model is a simplifying assumption that obeys the robustness of θ-parameter estimates to tree misspecification, reported by Yang, Goldman and Friday [130]. They declared that in estimating rate-matrix parameters "knowledge of the true phylogeny is not very important, as long as a sufficiently realistic model of evolution is adopted". In cases where DNA data are suspected to have undergone recombination of the type described in the

footnote in Section 5.2.2, an analysis with the $Q+t$ mixture model is not recommended. The reason is because contiguous subsequences of recombinant bacterial or viral DNA are expected to obey different phylogenetic trees. Therefore, a model that characterises the entire alignment by a single phylogenetic tree, such as the $Q+t$ mixture model, may be an inadequate description of this type of data.

By distinguishing evolutionary classes through specific Q-matrices and sets of branch lengths, the $Q+t$ mixture model is able to give a more complete description of the process underpining the evolution of DNA, unseen by a mixture of only Q-matrices like (5.3), or by the overall-rate model (5.1).

5.5.3 Possible extensions

The first extension for the $Q+t$ mixture model concerns dependence between observables. If we suppose that sites are generated according to the mixture model in (5.4), but now the allocation variables are the states of a Markov chain, then the allocation for site n is j with probability $p_{z_{n-1}j}$, which depends on the underlying allocation of the previous site, z_{n-1}. This is called a *hidden Markov model* and a thorough review of the topic can be found in [76, ch. 13]. Such an extension, however, would suffer from the problem of 'directionality' imposed by the Markov chain. The model would rely on adjacent sites in the observables actually having some direct effect on one another and possible biases may be introduced by the removal of sites resulting in consecutive observable sites not being direct neighbours in nature. Additionally, a hidden Markov formulation assumes that neighbouring sites are more likely to belong to the same component, which is a valid assumption for scenarios of recombination and insertion/deletion of adjacent sites. But this assumption is unrealistic when modelling the evolution of codon positions within a protein-coding region (Section 5.2.3). Some hidden Markov methods in phylogenetics include the model by Felsenstein and Churchill [28] presented above, the work done by Dirk Husmeier and his group [54, 53, 52], and the method by Webb, Hancock and Holmes [121].

An additional possible extension is to include multiple phylogenetic trees, which would result in greater breadth of the model. The computational cost of such an enhancement, however, is uncertain and so is its feasibility. We have not attempted such a formulation and this could be considered as an area for future work.

A more straightforward enhancement of the $Q+t$ mixture model would specify prior distributions for branch lengths with different values of β for different mixture components. That is, the prior distribution for the hth individual branch length of the jth mixture component could be specified as $t_{(h,j)} \sim Exp(\beta_j)$ independently for $h = 1, \ldots, 2S-3$ and $j = 1, \ldots, k$. This would allow different prior beliefs for different components to be included in the model.

5.5.4 Similar mixture models

In 2008, Pagel and Meade [89] modified their model (5.3) to allow for several sets of branch lengths along with multiple Q-matrices. They introduced a phylogenetic mixture similar in spirit to the $Q + t$ mixture model, but different in structure. Their model is a 'mixture of mixtures' that includes a dissimilar number of sets of lengths and Q-matrices. They did this as a way of reducing the computational cost, since they found cases in which less branch-length classes are required than rate-matrix classes. Under this model, the distribution for an observation x_n has the form:

$$x_n \sim \sum_{m=1}^{g} \upsilon_m \sum_{j=1}^{k} \omega_j p(x_n|\phi, \boldsymbol{t}_m, \boldsymbol{\theta}_j), \quad \text{independently for sites } n = 1, \ldots, N \qquad (5.11)$$

where $\upsilon_1, \ldots, \upsilon_g$ and $\omega_1, \ldots, \omega_k$ are mixture proportions corresponding to the different 'nested' mixtures. In addition to the structural differences between the $Q + t$ mixture and model (5.11), the main discrepancy lies in the interpretation. As opposed to Pagel and Meade's model, the phylogenetic mixture introduced in this thesis is formulated as a missing-data problem. Making simultaneous inference on both model parameters and allocation variables is one of the strengths of the $Q + t$ mixture. Inferring the allocation variables provides a natural framework for site classification and allows for a deeper understanding of the evolutionary process generating the data.

Also in 2008, Meade and Pagel [77] published a Bayesian mixture on (only) sets of branch lengths. This, as a means to account for heterotachy. In that same year, Kolaczkowski and Thornton [61] constructed a similar branch-length mixture but they estimated it within a maximum-likelihood framework. None of these authors, however, attempted a missing-data reformulation of their mixtures to allow for site classification.

Chapter 6

MCMC methods for the phylogenetic mixture model

6.1 Introduction

This chapter describes the MCMC methods used for estimating the $Q + t$ mixture model. The use of a missing-data reformulation allows the complicated structure of the mixture to be decomposed into simpler structures. As a result, the parameters of each component can be simulated conditionally on the allocations $\boldsymbol{z}$, taking into account only the sites that have been allocated to that component at a particular MCMC iteration. In most instances, the calculation of the acceptance probability of a Metropolis-Hastings step greatly simplifies as a direct consequence of this.

The chapter starts by presenting the moves used to traverse the space of mixture parameters. Most of these moves were already discussed in Chapter 4 in the context of the homogeneous phylogenetic model. The problem of the potential lack of identifiability of the mixture components, or *label-switching*, is briefly discussed in this chapter. If present, this phenomenon can affect the validity of inferences on component-specific quantities from MCMC output. Fortunately, label-switching does not always occur and if it does, it can be detected by visual inspection of the output.

Throughout the run, values of both $\boldsymbol{z}$ and model parameters are generated from their joint posterior distribution. Once samples of $\boldsymbol{z}$ have been obtained, they can be used to generate frequency counts of allocation of a site to a particular component. In the type of applications this thesis is concerned with, inference on allocation variables and the resulting frequency counts are of great explanatory interest as they provide us with the means for classifying the sites. This chapter discusses the details on such frequency counts and finally, it presents an overall discussion.

6.2 Moves for the $Q + t$ mixture model

The implementation of the MCMC sampler for the $Q + t$ mixture model involves the following move types:

(a) updating the mixture proportions $\boldsymbol{\omega}$;
(b) updating the phylogenetic tree ϕ;
(c) updating a branch length $t_{(h,j)}$, for $h = 1, \ldots, (2S-3)$, and $j = 1, \ldots, k$;
(d) updating the substitution rates $\boldsymbol{r}_j$, for $j = 1, \ldots, k$;
(e) updating the stationary probabilities $\boldsymbol{\pi}_j$, for $j = 1, \ldots, k$;
(f) updating the allocation z_n, for $n = 1, \ldots, N$.

One complete pass over these six moves will be called an *iteration* and is the basic time step of our algorithm. Therefore, a single iteration consists of a step to update the mixture proportions, a step to update the phylogenetic tree, $k(2S-3)$ steps to update all $(2S-3)$ individual branch lengths for each of the k mixture components, k steps to generate new substitution rates for all components, k steps to propose new stationary probabilities for all components, and N steps to propose new allocations for all sites.

Moves (b)-(f) are Metropolis-Hastings while move (a) is a Gibbs step. (Both the Metropolis-Hastings algorithm and the Gibbs sampler were discussed in Section 2.4.) Most of these moves have been thoroughly discussed in Chapter 4 already; the exceptions are (a) and (f). Details of moves (a) and (f) are now presented, together with a derivation of the corresponding acceptance probability, when pertinent. In the following, we use the notation $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k) = \big((\boldsymbol{t}_1, \boldsymbol{\theta}_1), \ldots, (\boldsymbol{t}_k, \boldsymbol{\theta}_k)\big)$ to denote the set of component-specific branch-length parameters and rate-matrix parameters and $\boldsymbol{z} = (z_1, \ldots, z_N)$ to denote the vector of allocation variables.

6.2.1 Updating the mixture proportions

A defective proposal

To update the vector of mixture proportions $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k)$, the algorithm samples from the full conditional posterior distribution for $\boldsymbol{\omega}$, given by

$$\begin{aligned} p(\boldsymbol{\omega}|\boldsymbol{z}, \phi, \boldsymbol{\psi}, \boldsymbol{x}) &\propto p(\boldsymbol{\omega}) p(\boldsymbol{z}|\boldsymbol{\omega}) \\ &\propto \prod_{j=1}^{k} \omega_j^{N_j} \end{aligned} \tag{6.1}$$

where $N_j = \sum_{n=1}^{N} I[z_n = j]$ is the sum of indicator functions and it amounts to the number of sites allocated to component j. The proportionality in the first line follows as we

only consider the factors in the joint distribution $p(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi}, \boldsymbol{x})$ that involve $\boldsymbol{\omega}$ (see equation (5.6)). The proportionality in the second line is because the prior distribution for $\boldsymbol{\omega}$ is $Dir_k(1, \ldots, 1)$ (Section 5.3.1). The product $\prod_{j=1}^{k} \omega_j^{N_j}$ follows from the choice of prior for an allocation variable (probability (5.8)). Notice that this product is the kernel of a Dirichlet distribution with parameters $(1 + N_1, \ldots, 1 + N_k)$. Thus a new vector of mixture proportions is generated by sampling from

$$\boldsymbol{\omega}' \sim Dir_k(1 + N_1, \ldots, 1 + N_k) \tag{6.2}$$

and candidate $\boldsymbol{\omega}'$ is always accepted as this is a Gibbs step. Proposal distribution (6.2) tells us that candidate mixture proportions will be generated according to the number of sites allocated to each component at a given MCMC iteration.

A difficulty of this updating mechanism is the existence of computational trapping states, or 'zero-stickiness'. When one of the components of the mixture is nearly empty, the chain becomes effectively trapped at the zero-boundary. This can be further investigated by deriving the variance of a candidate mixture proportion for component j, under proposal distribution (6.2):

$$\begin{aligned} Var(\omega_j') &= \frac{(1 + N_j)\,(\eta_0 - (1 + N_j))}{\eta_0^2\,(\eta_0 + 1)} \\ &= \frac{(1 + N_j)\,(k + N - 1 - N_j)}{(k + N)^2\,(k + N + 1)} \end{aligned}$$

where $\eta_0 = \sum_{j=1}^{k} 1 + N_j = k + N$. In the limit $N_j \to 0$,

$$Var(\omega_j') = \frac{k + N - 1}{(k + N)^2\,(k + N + 1)}$$

and for large N, this variance becomes $Var(\omega_j') \approx \frac{1}{(k+N)^2} \approx 0$. The phenomenon of the 'zero-stickiness' was previously encountered in this thesis (Sections 4.4.4 and 4.4.6). As before, we implement an alternative proposal in which the centre of the Dirichlet distribution is shifted by a quantity $\varepsilon > 0$. We believe that this is the most cost-effective way of resolving the problem.

An improved proposal

To generate candidate values for the vector of mixture proportions, the algorithm samples from an ε-corrected Dirichlet distribution of the form:

$$\boldsymbol{\omega}' \sim Dir_k(1 + N_1 + \varepsilon, \ldots, 1 + N_k + \varepsilon), \tag{6.3}$$

where $\varepsilon > 0$ is a small quantity that shifts the centre of the distribution. This approach resolves the problem since the algorithm is able to escape from states where one of the components is attributed very few (or none) sites.

6.2.2 Updating an allocation

Let z_n be the current allocation for site n. A candidate allocation z'_n is proposed by randomly drawing a value from the set $\{1, \ldots, k\} \setminus \{z_n\}$, where all elements in this set are equally likely to be drawn. The proposal distribution of the forward move is given by the probability of drawing a value from a set with $k-1$ elements (where each element is equally likely). This is, $q(z_n, z'_n) = \frac{1}{k-1}$. The reverse move has exactly the same proposal distribution and so the proposal ratio simplifies to $\frac{q(z'_n, z_n)}{q(z_n, z'_n)} = 1$. The probability of accepting the candidate allocation is:

$$\begin{aligned} \alpha(z_n, z'_n) &= min\left(1,\ \frac{p(z'_n|\boldsymbol{\omega})}{p(z_n|\boldsymbol{\omega})}\ \frac{L(\boldsymbol{z}', \phi, \boldsymbol{\psi}|\boldsymbol{x})}{L(\boldsymbol{z}, \phi, \boldsymbol{\psi}|\boldsymbol{x})}\ \cancelto{1}{\frac{q(z'_n, z_n)}{q(z_n, z'_n)}}\right) \\ &= min\left(1,\ \frac{\omega_{z'_n}}{\omega_{z_n}}\ \frac{L(\phi, \boldsymbol{t}_{z'_n}, \boldsymbol{\theta}_{z'_n}|x_n)}{L(\phi, \boldsymbol{t}_{z_n}, \boldsymbol{\theta}_{z_n}|x_n)}\right) \end{aligned} \tag{6.4}$$

where the prior ratio $\frac{p(z'_n|\boldsymbol{\omega})}{p(z_n|\boldsymbol{\omega})}$ is derived from probability (5.8). Since only the allocation for site n is changing, the ratio $\frac{L(\boldsymbol{z}', \phi, \boldsymbol{\psi}|\boldsymbol{x})}{L(\boldsymbol{z}, \phi, \boldsymbol{\psi}|\boldsymbol{x})}$ simplifies to the quotient of the likelihood for site n given allocation z'_n over the likelihood for this same site given z_n.

6.3 Label switching

The application of MCMC methods to the analysis of mixtures, as convenient as it may seem, presents some practical difficulties. One of such difficulties is the *non-identifiability* of the components under prior distributions that do not contain any information regarding which component should be labelled what. The components of a mixture model are non-identifiable since the same likelihood function is obtained for any permutation of the component labels. This is the so called *label switching* problem [56].

In order to illustrate this, consider data $x_1, \ldots, x_N$, assumed to be independent observations from the distribution

$$x_n \sim \frac{1}{2}\, N(\mu_1, 1) + \frac{1}{2}\, N(\mu_2, 1)$$

where $N(\mu, 1)$ is the univariate normal density with mean μ and known variance 1. Suppose that at certain iteration of an MCMC sampler, a point $(\mu_1, \mu_2) = (0, 1.5)$ is visited so that the likelihood function, evaluated at this point, is:

$$\prod_{n=1}^{N}\left(\frac{0.5}{\sqrt{2\pi}}e^{-\frac{1}{2}x_n^2}+\frac{0.5}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_n-1.5)^2}\right) \tag{6.5}$$

Suppose that, at a different iteration, the sampler visits $(\mu_1, \mu_2) = (1.5, 0)$. The evaluation of the likelihood at this new point is:

$$\prod_{n=1}^{N}\left(\frac{0.5}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_n-1.5)^2}+\frac{0.5}{\sqrt{2\pi}}e^{-\frac{1}{2}x_n^2}\right) \tag{6.6}$$

Expressions (6.5) and (6.6) are identical, and the likelihood is not affected by μ_1 being labelled 'one' and μ_2 'two', or the other way round.

Suppose further that the posterior distribution has a mode at $(0, 1.5)$. Given that the likelihood function has the same value either if $(\mu_1, \mu_2) = (0, 1.5)$ or if $(\mu_1, \mu_2) = (1.5, 0)$, the posterior distribution effectively has two symmetric modes, one at each of these points. A sampler may therefore be attracted to these two high-density regions, in which case, the MCMC output will show several jumps between the two posterior symmetric modes.

It is invalid to draw inference of component-specific quantities whenever the MCMC output has been masked by label switching. Fortunately the problem can be detected by visual inspection; if a switch in values for the allocation variables is observed, one can suspect of the presence of label switching during simulation. If present, label switching can be resolved by one of a number of methods suggested in the literature. For example, an informative prior distribution could provide a rule for labelling the mixture components; one could impose a prior in which the means of the components are in increasing numerical order or the mixture proportions in non-decreasing order. If the label-switching problem is not solved at the simulation stage, one could still use a number of existing alternatives to fix it at the inferential stage (e.g. [18, 11]).

As suggested by this example, the problem is typically found when the components of the mixture are not far apart. In this mixture for instance, the densities $N(0, 1)$ and $N(1.5, 1)$ overlap each other and it is easy for the sampler to switch the labels that identify the components during simulation. In the analyses carried out as part of this thesis, the MCMC output was checked for label-switching. This problem was not encountered perhaps due to the components of the mixture being well distinguished in all cases. The lack of label-switching enabled us to make valid inferences of quantities of interest without the need to resolve this problem.

6.4 Inference on allocation variables

The method presented in Section 6.2 simulates a Markov chain of mixture parameters and a chain of allocation variables from the joint posterior distribution $p(\boldsymbol{\omega}, \boldsymbol{z}, \phi, \boldsymbol{\psi}|\boldsymbol{x})$. Once the chains have been produced and checked for convergence to stationarity, good mixing and lack of label-switching, they can be used to make reliable inferences about the posterior distribution. In this section, we consider inferences from the chain of allocation variables.

Consider a sample $z_n^{(1)}, \ldots, z_n^{(M)}$ of the allocation for site n, generated from an MCMC run of length M after burn-in. Variable $z_n^{(i)}$ indicates the identity of the component to which site n is allocated at iteration i and it takes values in the set $\{1, \ldots, k\}$. The sample $z_n^{(1)}, \ldots, z_n^{(M)}$ can be used to count the number of times that site n was allocated to component j throughout the run. This frequency count, divided by the total number of samples, M, gives the *posterior classification probability* of site n to component j.

Additionally, we can pick the component with the highest posterior classification probability as the single 'most likely' component for site n, given the data and the model. This gives a criterion for *optimal classification* for site n.

6.5 Discussion

Most of the updating mechanisms presented in this chapter have been thoroughly discussed in Chapter 4. The exceptions are the proposals for the mixture proportions and the allocation variable. The former corresponds to a widely-used mechanism, which generates transitions for $\boldsymbol{\omega}$ by sampling from its full conditional posterior distribution (see for example [96], [76, ch. 4]). The proposal for an allocation, on the other hand, is particular to this thesis and we now discuss it.

In the MCMC literature, it is common to find a proposal mechanism for an allocation z_n that differs from the one discussed in Section 6.2.2. A more conventional way of generating transitions for z_n is by sampling from its full conditional posterior distribution, given by

$$p(z_n = j|\boldsymbol{\omega}, \phi, \boldsymbol{\psi}, x_n) = \frac{\omega_j \, L(\phi, \boldsymbol{t_j}, \boldsymbol{\theta_j}|x_n)}{\sum_{i=1}^{k} \omega_i \, L(\phi, \boldsymbol{t_i}, \boldsymbol{\theta_i}|x_n)}, \tag{6.7}$$

for $j = 1, \ldots, k$. Sampling from this distribution makes the move into a Gibbs step. In our algorithm, however, this move is a Metropolis-Hastings step, which arguably entails greater complexity than its Gibbs equivalent. Nevertheless, the cost of sampling from the full conditional (6.7) in a phylogenetic context can be high. The denominator in (6.7) requires us to evaluate the likelihood at all possible allocations $1, \ldots, k$. In Section 3.6 we discussed the intrinsic complexity of calculating the likelihood function for phylogenetic

parameters. This calculation required to sum over all possible assignments of states at the interior nodes of the tree and for large analyses, the repeated evaluation of the likelihood at all possible allocations can be computationally prohibitive.

When designing our MCMC sampler, we realised the difficulties of sampling from the full conditional for z_n. It was then when we decided to update this variable via the computationally-cheaper mechanism presented in Section 6.2.2. This has given fine results and is evidence of the great flexibility of MCMC methods; one can adopt any proposal scheme as long as aperiodicity and irreducibility of the Markov chain are ensured and the proposal can be easily sampled and evaluated. We must point out, however, that the computational cost of both proposals, the one from Section 6.2.2 and the proposal in (6.7), is comparable when the number of components in the mixture is two. This is because the acceptance probability of our proposal always requires the evaluation of the likelihood function at two distinct components (acceptance probability (6.4)).

In spite of its potentially higher computational cost, generating transitions for z_n through (6.7) might result in nice mixing properties. However, that might also be the case for alternative Metropolis-Hastings proposals. The breadth of Metropolis-Hastings MCMC is such that it is always possible to design new transition mechanisms as long as irreducibility holds. We have not investigated the properties of other transition mechanisms for z_n and there is scope for this to be pursued as future research.

Chapter 7

Analysis of the mitochondrial DNA of primates

7.1 Introduction

This chapter assesses the $Q+t$ mixture model by application to real DNA data. The 'primate mtDNA alignment', as we will refer to it, has been extensively used to validate proposed methodologies in the past (e.g. [125, 124, 130, 126, 131, 63, 112, 111]). It is a well understood dataset within the phylogenetics community and this makes it suitable for validating our $Q+t$ mixture model. Previously published analyses of these data have made a number of simplifying assumptions for tractability, including the prior partitioning of the alignment into evolutionary classes and the inclusion of only a moderate number of sequences when classification of the sites is one aim of the analysis. We are not aware of published methodologies that infer the classification structure directly from the primate mtDNA alignment while including a reasonable number of DNA sequences. In this chapter, we demonstrate that the $Q+t$ mixture is able to handle both.

7.2 A close look at monomorphic sites

Before we present the analysis of the primate mtDNA data, it is necessary to discuss the nature of a monomorphic site and the effect that this type of data have in phylogenetic analyses. A *monomorphic site* is one in which all taxa share the same nucleotide character. For example, in a DNA alignment of four taxa, sites $x_n = (A, A, A, A)^T$ and $x_m = (G, G, G, G)^T$ are both monomorphic. Consider site $x_n = (A, A, A, A)^T$. The likelihood method reconstructs the history of these characters depending upon a phylogenetic tree, the length of the branches of the tree and the Markov process of nucleotide substitution assumed. In this example, consider the tree in Figure 7-1, denoted by ϕ, and assume, without loss of generality, that the characters evolve under the simple JC model (Section 3.3.2). Later in this section we will show, through a simulated example, that the results that we now discuss in the context of the JC model also hold when assuming the GTR model.

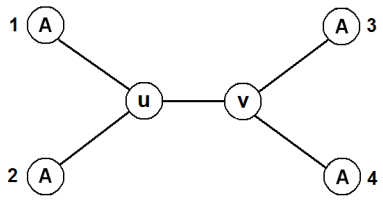


Figure 7-1: A simple phylogenetic tree of four taxa with the same monomorphic states at the leaves.

7.2.1 Branch lengths that approach zero

In order to compute the likelihood function, assume that the phylogenetic tree ϕ, in Figure 7-1, is rooted at node 1. The probability of observing site $x_n = (A, A, A, A)^T$ given tree ϕ, a set of branches with total length h and the JC model of character substitution is calculated as (Section 3.6):

$$p(x_n|\phi, h) = \sum_{u\in\mathcal{I}}\sum_{v\in\mathcal{I}} \pi_A\, p_{Au}\left(\frac{h}{5}\right) p_{uA}\left(\frac{h}{5}\right) p_{uv}\left(\frac{h}{5}\right) p_{vA}\left(\frac{h}{5}\right) p_{vA}\left(\frac{h}{5}\right), \qquad (7.1)$$

where $\mathcal{I} = \{A, C, G, T\}$ is the set of DNA characters. (The total branch length h has been arbitrarily divided by 5 because this is the number of branches in a four-leaf tree and it is not relevant for this exercise to talk about individual branch lengths.) Here, the JC transition probabilities are given by:

$$\begin{aligned} p_{ii}(t) &= \frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}t} \\ p_{ij}(t) &= \frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}t} \end{aligned} \qquad (7.2)$$

for $t \geq 0$. These probabilities have been standardised so that the expected total rate of substitution is one (Section 3.4.2). Table 7.1 shows probability (7.1) evaluated at different total branch lengths. Notice how the likelihood increases as the value for h decreases. This is hardly surprising as a monomorphic site, with no variation in characters among taxa, will favour a tree reconstruction with short branches, agreeing with very few expected character substitutions.

After expanding the summations in (7.1) and some algebraic manipulation, it can be shown that the likelihood becomes:

$$p(x_n|\phi, h) = \frac{1}{4}\quad p_{ii}\left(\frac{h}{5}\right)^5 + 6p_{ii}\left(\frac{h}{5}\right)^2 p_{ij}\left(\frac{h}{5}\right)^3 + 3p_{ii}\left(\frac{h}{5}\right)\left(p_{ij}\left(\frac{h}{5}\right)^4 + 6p_{ij}\left(\frac{h}{5}\right)^5\right)\left(\qquad (7.3)$$

where $p_{ii}(t)$ and $p_{ij}(t)$ are given by (7.2), and the JC stationary probability for character A

total length h	$p(x\|\phi, h)$
0.00001	0.24999
0.001	0.24975
0.1	0.22628
1.0	0.09543
10.0	0.00403

Table 7.1: Probability of a monomorphic site $x = (A, A, A, A)^T$, given the JC model, the tree in Figure 7-1 and a total branch length indicated in the first column.

is $\pi_A = \frac{1}{4}$. When the total branch length is small, the term $\frac{1}{4}\, p_{ii}(\frac{h}{5})^5$ is the only one that plays an important part in expression (7.3). For instance, when $h = 0.1$, this term accounts for 99.99% of the total probability $p(x_n|\phi, h)$. Since this term corresponds to an scenario in which ancestral characters at nodes u and v are both 'A', we can say that for small h, probability $p(x_n|\phi, h)$ has most of its weight on the tree reconstruction that involves no character changes at all. As the total branch length increases (indicating a greater expected number of character substitutions), the percentage probability placed at the tree reconstruction involving no character changes decreases. For instance, when $h = 1.0$, the weight of term $\frac{1}{4}\, p_{ii}(h/5)^5$ is 99.77% whereas when $h = 10.0$, this term accounts for only 15.58% of the total probability $p(x_n|\phi, h)$.

From Table 7.1, the most likely tree reconstruction for a monomorphic site is one with very little evolutionary divergence, i.e. a tree with very short branches. If the branches are short, a scenario involving no character changes (i.e. a tree where the characters at the interior nodes are equal to the observed characters at the leaves) is favoured.

7.2.2 Unresolved tree topology

Phylogenetic trees with interior branch lengths that approach zero are usually called *unresolved* or *star trees*. Consider the tree in Figure 7-1, this tree has only one interior branch. In the limit where the length of this interior branch approaches zero, the tree resembles more and more four rays radiating from a single central point, hence the name 'star'. A star-tree indicates that the data lack phylogenetic signal to *resolve* which two nodes are closest neighbours. The resulting tree is a 'star' where all four leaves are closest neighbours to one another. The shorter the interior branches, the more unresolved (or star-like) the tree is.

A monomorphic site favours very short branch lengths and, consequently, a star-like tree that provides unresolved information about the evolutionary relatedness of a group of taxa. This is the reason why monomorphic sites are usually referred to as *phylogenetically uninformative*.

tuning parameter	*value*
δ	1.30
σ	0.05
α_1	700.00
α_2	600.00
ε_θ	0.0001

Table 7.2: Tuning parameters for the MCMC update mechanisms used during the analysis of the monomorphic alignment. Parameter δ tunes the BLM move; σ tunes the BLNA move; α_1 and α_2 tune the εDP moves for substitution rates and stationary probabilities, respectively and ε_θ is the size of the correction for the εDP move.

7.2.3 A simulated example

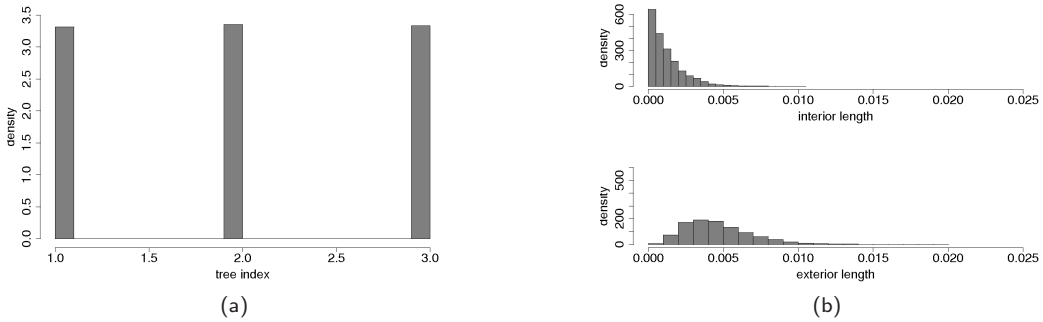
We now show that the results from the previous section, based on the JC model, also hold when assuming the GTR model. Consider a synthetic alignment of size 4×800 containing only monomorphic sites. We fit these data with the homogeneous phylogenetic model and assume a GTR process of character substitution. We set the hyperparameter of the exponential prior on a branch length to $\beta = 10$ (prior distribution (3.21)). The MCMC sampler from Section 4.4 was run for $50\,000$ iterations and the initial $12\,500$ steps were discarded as burn-in. The tuning parameters of the proposals were set to the values shown in Table 7.2.

Figure 7-2(a) shows the histogram of the sampled tree topologies. This histogram is evidence of the unresolved tree topology in a purely monomorphic alignment. All trees in the tree space have equal posterior support; no single tree provides a better explanation of the monomorphic data than any other.

Figure 7-2(b) shows the histogram of the sampled branch lengths, with an ergodic average of 0.0012 for the interior branch and 0.0046 for the sum of the four exterior branch lengths. As expected, the monomorphic data support branch lengths that lie very close to zero, indicating a negligible number of substitution events.

In the GTR model, the Q-matrix is parametrised in terms of the substitution rates $\boldsymbol{r} = (r_{AC}, r_{AG}, r_{AT}, r_{CG}, r_{CT}, r_{GT})$ and the stationary probabilities $\boldsymbol{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T)$ (Section 3.3.5). Figure 7-3 shows the histograms of sampled substitution rates. The ergodic averages for all rates is close to $\frac{1}{6}$, which is not surprising since monomorphic data have undergone little or no substitutions, which pulls the rate parameters toward zero. (Our model constrains the six rates of substitution to add up to one, hence the ergodic averages for $r_{AC}, \ldots, r_{GT}$ are localised above $\frac{1}{6}$.) The histograms of sampled stationary probabilities are shown in Figure 7-4. The ergodic averages for π_A, π_C, π_G and π_T are localised above the true proportion of observed 'A's, 'C's, 'G's, and 'T's in the alignment.

In summary, the identical-character composition of a monomorphic site favours tree reconstructions with very short branches. Interior branches that approach zero result in



(a) (b)

Figure 7-2: (a) Histograms of posterior phylogenetic tree and (b) interior/exterior branch lengths for the monomorphic alignment. The results correspond to $37\,500$ iterations, after a burn-in period of $12\,500$ iterations.

star-like trees that provide unresolved information about the evolutionary relatedness among taxa. An alignment that contains both monomorphic and non-monomorphic (or polymorphic) sites is a compromise between 'non-evolving' and 'evolving' positions. Monomorphic positions will pull branch lengths towards zero while all other sites may favour tree reconstructions with longer branches. Given that polymorphic positions tend to follow less ambiguous patterns of evolution than their monomorphic counterparts, we sometimes adopted the approach of removing all monomorphic positions and analysing only polymorphic ones. When removing monomorphic sites from an alignment, we will make it explicit and, otherwise, it should be assumed that the alignment has not been modified in any form. Thinning out monomorphic sites introduces bias to parameter estimates as we are implicitly assuming that a polymorphic alignment is exclusively formed by polymorphic sites (whereas, in reality, some monomorphic observations may occur by chance). In the following section, we discuss some of the consequences of removing all monomorphic positions from an alignment.

7.2.4 Some consequences of thinning out monomorphic sites

By removing all monomorphic positions from an alignment we are neglecting the fact that some 'polymorphic' sites may end up looking monomorphic purely by chance or by experiencing 'silent' substitutions that do not show in the data. As a result, parameter estimates from a purely polymorphic alignment may be biased if truly-occurring monomorphic sites are discarded. The parameter estimates that we obtain from a completely polymorphic alignment are, in effect, conditioned on the sites that we keep being truly polymorphic. Or, in mathematical notation:

Let $A(x_n)$ be the event that a site x_n is not monomorphic. Then, the likelihood of parameter θ conditioned on this event is written as

$$p(x_n|\theta;\, A(x_n)) = \frac{p(x_n, A(x_n)|\theta)}{p(A(x_n)|\theta)},$$

where $p(x_n, A(x_n)|\theta) = 0$ if x_n is a monomorphic site and $p(x_n, A(x_n)|\theta) = p(x_n|\theta)$ if x_n

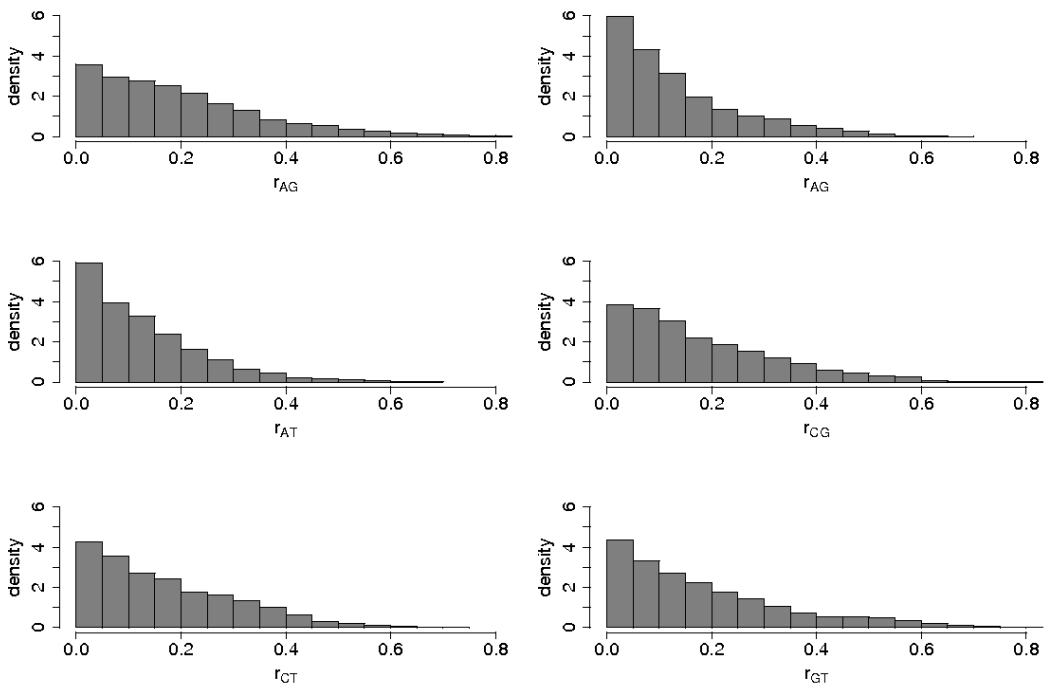


Figure 7-3: Histograms of sampled substitution rates for the monomorphic alignment. Monomorphic data have not undergone apparent character substitution and, therefore, do not contain information about the rates of substitution.

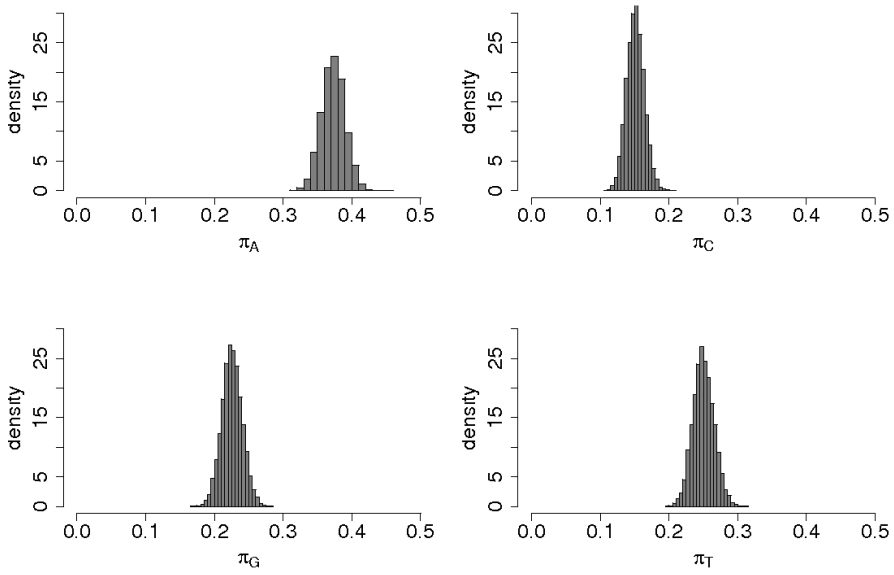


Figure 7-4: Histograms of sampled stationary probabilities for the monomorphic alignment. The ergodic average for stationary probability π_i is localised above the proportion of observed characters of type i in the alignment.

is not monomorphic. As we are interested in removing all monomorphic sites to end up with a completely polymorphic alignment, we have:

$$p(x_n|\theta;\, A(x_n)) = \frac{p(x_n|\theta)}{p(A(x_n)|\theta)}. \tag{7.4}$$

Calculating the 'biased' likelihood $p(x_n|\theta)$ is straightforward (Section 3.3.1) but calculating the correction factor $p(A(x_n)|\theta)$ requires more care. Factor $p(A(x_n)|\theta)$ represents the probability of x_n being non-monomorphic given θ. This probability can be estimated by simulating over a likely θ and counting the number of times that we end up with a polymorphic observation. Thus, we could for example obtain an estimate $\hat{\theta}$ based on a purely polymorphic alignment, generate a synthetic dataset using $\hat{\theta}$, and observe the proportion of polymorphic sites that occur in this synthetic alignment. Alternatively, we could calculate $p(A(x_n)|\theta)$ directly by considering the complementary event; we would then have to calculate the likelihood for the only four cases in which a site is monomorphic (i.e. A, C, G and T monomorphic sites).

7.3 The *primate mitochondrial DNA* alignment

7.3.1 The data

All living organisms are made up of cells. The cells of eukaryotic organisms (i.e. animals, plants, fungi and protists) have a well defined nucleus and contain specialised structures that perform specific functions, called *organelles*. One of these organelles, the *mitochondrion*, is a rod-shaped structure whose main function is to provide the cell with energy. Although most of the cell's DNA is packaged within the nucleus, mitochondria also have their own DNA. This genetic material, known as the *mitochondrial DNA* (mtDNA), is of particular interest to scientists because certain human disorders, such as cancer or Leber's optic atrophy, are known to be related to mutations in the mtDNA [80, 109].

In 1982, Brown, Prager, Wang and Wilson [7] obtained the sequences of a segment of mtDNA from five primates: human; gorilla; chimpanzee; orangutan; and gibbon. Their main interest was in studying the dynamics of the evolution of mtDNA by assessing the rate at which it experiences substitutions. A few years later, Hayasaka, Gojobori and Horai [48] published the DNA sequences for the same mitochondrial region from seven other species of primates. In 1995, Yang [126] combined the original five mtDNA sequences with four others extracted from the study by Hayasaka et al., to form a dataset containing a total of nine species: human; gorilla; chimpanzee; orangutan; gibbon; crab-eating macaque; common squirrel monkey; Philippine tarsier and ring-tailed lemur. This is the set of sequences that the phylogenetics community usually refers to as the *primate mtDNA* alignment.

The *primate mtDNA* alignment has size 9×888, with 43.69% of its sites monomorphic. It is formed by a transfer-RNA region (tRNA) and segments of the coding regions

for two proteins: protein ND4 and protein ND5. Transfer RNA is a small molecule in charge of translating the information encoded by the DNA into the protein alphabet. A fascinating aspect of this translating molecule is that all living organisms share a similar tRNA structure, which in evolutionary terms indicates that tRNA molecules have been highly conserved through millions of years of evolution. The protein-coding regions, on the other hand, are arrangements of non-overlapping groups of three nucleotides, called *codons*, that constitute the protein-alphabet (Section 3.3.3). The positions within a codon are referred to as the first, second and third codon positions. Thus, the protein-coding sequence $CGAACATTACCC$ has four codons: CGA, ACA, TTA and CCC, and the nucleotides occurring at the second codon position are: G, C, T and C.

Nucleotides at the second codon position ($cp2$) are known to undergo substitutions at a slow rate and, therefore, be highly conserved through evolution (Section 5.2). Nucleotides at the first and third codon positions ($cp1$ and $cp3$, respectively) experience substitutions at higher rates. This is related to the fact that a change in a nucleotide occurring at the third codon position does not always affect the resulting protein, but a change in a nucleotide at the second codon position may alter the final product with a higher probability and be potentially disastrous for an organism (Section 3.3.3). In terms of a DNA alignment, sites occurring at third codon positions are the most variable and are rarely monomorphic sites (only 8.65% of $cp3$-sites in the primate mtDNA alignment are monomorphic). Sites occupying the second codon position or belonging to the tRNA region are the most conserved ones (61.47% of $cp2$-sites are monomorphic and 62.37% of the sites at the tRNA region are monomorphic). Finally, 45.25% of $cp1$-sites are monomorphic. Incidentally, of the few monomorphic sites occurring at $cp3$, most of them are of the A type and only very few of the C type; none of them are monomorphic G or T sites.

For presentational purposes, we rearranged the *primate mtDNA* alignment so that sites occurring at the same codon position were next to one another, yielding an alignment of the form $|cp1|cp2|cp3|tRNA|$. The first three intervals correspond to codon positions one (sites $1-232$), two (sites $233-463$) and three (sites $464-694$), respectively, and the fourth interval corresponds to the tRNA region (sites $695-888$). If the sites form clusters according to their codon-position or tRNA nature as previously published studies have assumed (e.g. [126, 63]), a rearrangement of the form $|cp1|cp2|cp3|tRNA|$ will help displaying the classification probabilities in a neater way than an arrangement where the codon positions are scattered all across the alignment. Nevertheless, there is nothing in the formulation of the $Q+t$ mixture model that requires such a rearrangement; this is in contrast with the change-point method by Suchard, Weiss, Dorman and Sinsheimer [111] or certain applications of hidden Markov models (Section 5.2.3), which rely on correlated sites being close to one another.

7.3.2 The scientific question

We are interested in detecting evolutionary heterogeneity in the *primate mtDNA* alignment. If sites at the different codon positions and at the tRNA region contain heterogeneous evolutionary signals, we would expect them to be classified to different mixture components. However, we do not assume *a priori* that the heterogeneity in this alignment is due to the difference in codon positions and tRNA, but allow for alternative explanations to be inferred from the data. Neither we assume that the data are explained by a mixture of four components (corresponding to the three codon positions plus tRNA) but investigate the good fit of two, three and a four-component mixtures.

The scientific question may be stated as follows: *Is there any evolutionary heterogeneity in the primate mtDNA alignment and, if so, what is its nature?*

7.3.3 Previous analyses

In [126], Ziheng Yang accounted for rate heterogeneity among sites by partitioning the mtDNA alignment into four rate classes; three for the different codon positions and a fourth one for the tRNA region. He assumed that a site was *a priori* known to belong to one of the four classes. A site known to belong to class 1, the first codon position, was assigned an overall-rate parameter γ_1. A site known to belong to class 2, the second codon position, was given an overall-rate γ_2, and similarly for the remaining two classes; the third codon position (class 3) and tRNA region (class 4). He estimated the overall-rates for the different classes to be in the ratios $\gamma_1 : \hat{\gamma_2} : \hat{\gamma_3} : \hat{\gamma_4} = 1.00 : 0.47 : 3.24 : 0.58$. These results agree with the theory, since codon position three has the highest rate estimate. Notice the similarity between $\hat{\gamma_2}$ and $\hat{\gamma_4}$, the estimates for codon position two and the tRNA region, respectively. Both of them are known to be highly conserved and to undergo very few substitutions, which is captured by the estimated values. Yang later stated that these estimates were inaccurate due to some sites in the tRNA segment being *a priori* misclassified (mtprim9.nuc file in PAML [129]). This is an example of how prior classification of sites to evolutionary classes may be restrictive, as opposed to inferring the classification structure directly from the data.

When Larget and Simon [63] published the LOCAL mechanism for updating tree topologies (Section 4.3.2), they also studied the *primate mtDNA* dataset. Similar to Yang's approach, they assigned a label $z \in \{1, 2, 3, 4\}$ to each site, based on prior knowledge of the site membership to one of the four classes: $cp1$, $cp2$, $cp3$ or tRNA. In their analysis they obtained estimates that agreed with the evolution for different codon positions and for the highly-conserved tRNA. Their approach relied on prior knowledge of the membership of a site to an evolutionary class and so is prone to errors. In contrast to Yang's and Larget and Simon's methods, the $Q + t$ mixture model does not assume any prior knowledge about the membership of a site to a class but rather lets the data speak for themselves.

In [111], Suchard, Weiss, Dorman and Sinsheimer analysed a four-sequence version of the *primate mtDNA* alignment as a means to validate their change-point model. The sites were manually rearranged to create four successive groups that made the alignment physically look like $|cp1|cp2|cp3|tRNA|$, where the sites that belong to the first codon position were gathered together to the left, sites originating from the second codon position were placed next, and so on. Their approach relies on the alignment being rearranged in this way so that correlated sites are close to one another. Contrary to other approaches, this method does not assume that site membership is *a priori* known. Their method adequately infers the locations at which one codon position ends and a new one begins. Their estimates were consistent with the notion of $cp3$ experiencing substitutions at a high rate, and with $cp2$ and tRNA undergoing substitutions at low rates. However, their method only works for a rather modest number of DNA sequences; Suchard and co-workers originally studied an alignment of four sequences and later, in [79], they extended the model to handle up to eight sequences.

7.4 A two-component analysis

We fitted the primate mtDNA data with a two-component $Q+t$ mixture model. Figures 7-6, 7-7 and 7-8 summarise the results corresponding to $30\,000$ iterations, after a burn-in of $10\,000$ steps, based on the MCMC sampler in Section 6.2. The hyperparameter of the exponential prior distribution on a branch length was set to $\beta = 10$ (in (3.21)). Preliminary runs indicated that other choices of β yielded similar results. The tuning parameters for the MCMC moves are given in Table 7.3. A number of initial exploratory runs were performed to help choosing these values, which were selected in order to achieve good mixing of the chain. The initial runs also helped to monitor convergence to stationarity.

The estimated posterior classification probabilities to component 2, i.e. $p(z_1 = 2|\boldsymbol{x})$, $p(z_2 = 2|\boldsymbol{x})$, $\ldots$, $p(z_{888} = 2|\boldsymbol{x})$, are shown in Figure 7-5. The boundaries of the three

tuning parameter	*value*
δ	1.50
σ	0.06
α_1	800.00
α_2	600.00
ε_θ	0.0001
ε_ω	0.0001

Table 7.3: Tuning parameters for the MCMC update mechanisms used during the analysis of the *primate mtDNA* alignment. Parameter δ tunes the BLM move; σ tunes the BLNA move; α_1 and α_2 tune the εDP moves for substitution rates and stationary probabilities, respectively; ε_θ is the size of the correction for the εDP move and ε_ω is the correction for the mechanism updating the mixture proportions.

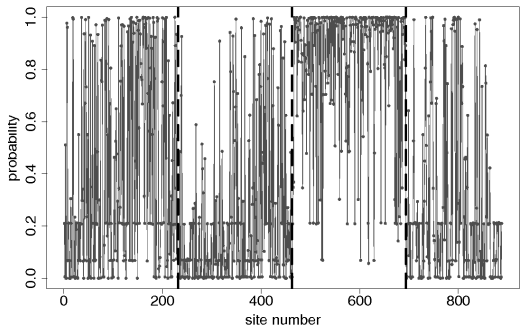


Figure 7-5: Estimated posterior classification probabilities to the *polymorphic component* for the *primate mtDNA* alignment. Results obtained from an analysis with a two-component $Q + t$ mixture. The boundaries between the three codon positions and the tRNA region have been marked as dotted lines to resemble the arrangement $|cp1|cp2|cp3|tRNA|$.

codon positions and the tRNA region are marked as dotted lines to resemble the arrangement $|cp1|cp2|cp3|tRNA|$. By inspecting the classification probabilities and the mtDNA alignment we find a correspondence between high probabilities of allocation to component 2 and polymorphic sites. In other words, sites with a high probability in Figure 7-5 mostly correspond to polymorphic positions; therefore, we refer to component 2 as the *polymorphic component*. In contrast, high probabilities of allocation to component 1 coincide with a combination of both polymorphic and monomorphic sites. However, it is monomorphic sites that prevail and so we refer to component 1 as the *monomorphic component*.

The effect of horizontal lines in Figure 7-5 is due to the presence of monomorphic sites in the alignment. A monomorphic site of one type is indistinguishable from another monomorphic site of the same type in terms of the likelihood function. Consequently, all monomorphic positions of one type have the same posterior classification probability to a given component. To illustrate this, let us think of the posterior probability for z_n as the prior beliefs about z_n updated by the observed x_n. Then, given an observed site $x_n = (A, A, A, A)$ and a prior $p(z_n = j|\boldsymbol{\omega}) = \omega_j$, the marginal posterior probability $p(z_n = j|x_n)$ will be exactly the same as that for an observed site $x_m = (A, A, A, A)$ with prior $p(z_m = j|\boldsymbol{\omega}) = \omega_j$. Moreover, to completely convince ourselves this is the case, it is possible to explicitly write the marginal probability $p(z_n = j|x_n)$ as

$$p(z_n = j|x_n) = \frac{p(z_n = j|\boldsymbol{\omega})\, p(x_n|\phi, \boldsymbol{t}_j, \boldsymbol{\theta}_j)}{\sum_{i=1}^{k} p(z_n = i|\boldsymbol{\omega})\, p(x_n|\phi, \boldsymbol{t}_i, \boldsymbol{\theta}_i)}$$

and for identical observations $x_n = x_m$, this posterior probability is indistinguishable.

A two-component analysis captures the evolutionary differences between the well conserved regions ($cp2$ and tRNA) and the highly-variable $cp3$ region. This distinction is based

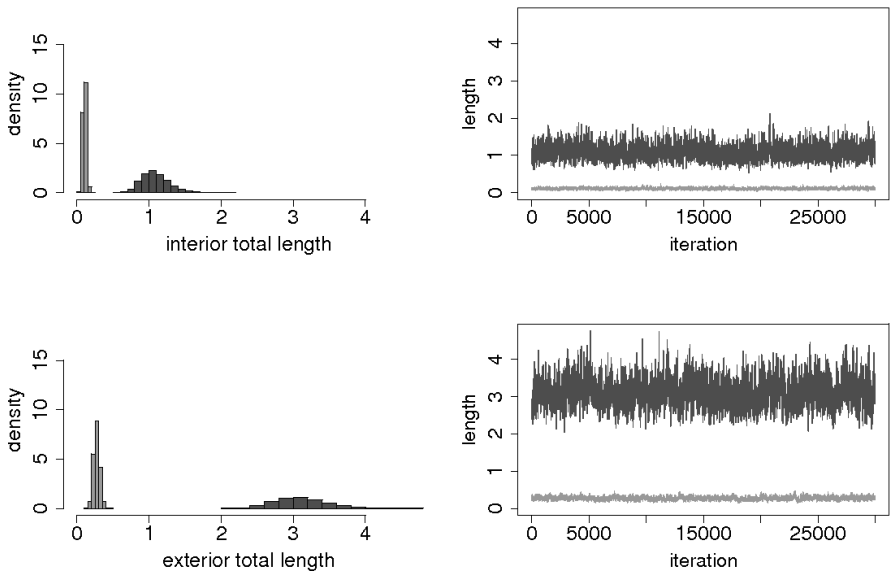


Figure 7-6: On the left-hand side, histograms of the sampled total length of interior and exterior branches for the *primate mtDNA* alignment. The traces of sampled values are shown on the right-hand side of each histogram. The component in light grey corresponds to the *monomorphic component*.

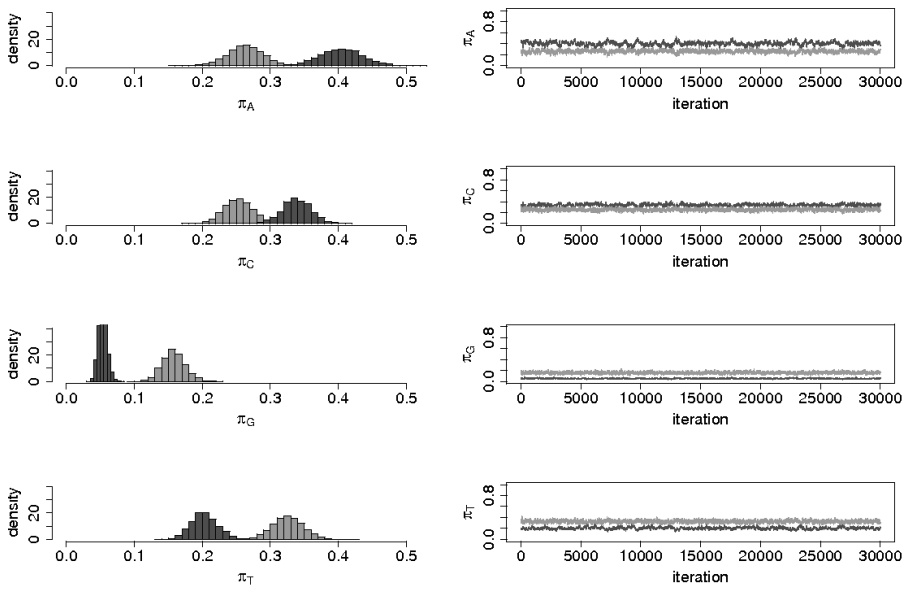


Figure 7-7: Histograms of sampled stationary probabilities for the *primate mtDNA* alignment. The traces of sampled values are shown to the right of each histogram. The component in light grey corresponds to the *monomorphic component*.

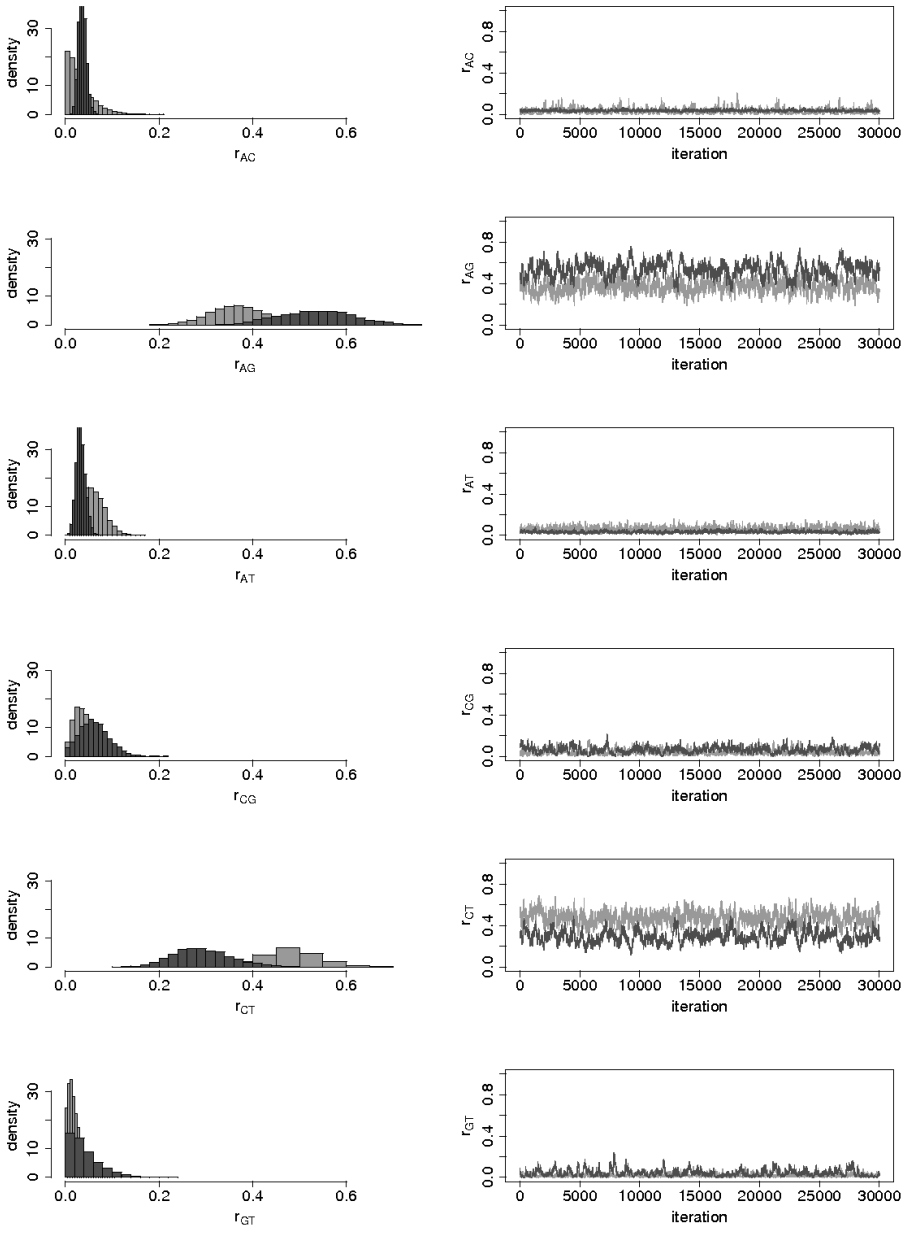


Figure 7-8: Histograms of posterior substitution rates for the *primate mtDNA* alignment. The traces of sampled values are shown to the right of each histogram. The component in light grey corresponds to the *monomorphic component*.

on the proportion of monomorphic sites contained in the different codon positions and tRNA. A well-conserved region contains a high proportion of monomorphic sites, whereas a variable region contains few monomorphic sites. In Figure 7-5, there is a noticeable difference between the $cp2$ and $cp3$ intervals because $cp2$ is highly-monomorphic (and most of the sites within the $cp2$ region have a high probability of classification to the monomorphic component) but $cp3$ is highly-polymorphic. However, not only polymorphic sites occurring at $cp3$ are allocated to the polymorphic component; polymorphic sites all across the alignment have a high probability of classification to component 2. The evolutionary heterogeneity detected in the primate mtDNA alignment by a two-component $Q + t$ mixture is due to the differences between monomorphic and polymorphic sites and not to the differences between sites originating from different codon positions or the tRNA region.

Once the two mixture components have been associated with a monomorphic and a polymorphic class, it is natural to expect that the monomorphic class will favour tree reconstructions with branch lengths that approach zero (Section 7.2). In Figure 7-6, the component in light grey corresponds to the monomorphic component. This component displays very short total length for both interior and exterior branches, which agrees with our expectations.

Table 7.4 summarises the ergodic averages of model parameters and the estimated integrated autocorrelation times for the monomorphic ($j = 1$) and the polymorphic ($j = 2$) components. The ergodic averages for the rates of substitution agree with the bias that favours transitions (a substitution from $A \rightarrow G$ or $C \rightarrow T$) over transversions (any other possible substitution; Section 3.3.3). The polymorphic component has a low proportion of G characters, which is exhibited in the low value for $\bar{\pi}_{(G,2)}$. We do not know the biological reason for this.

In Table 7.4, we can observe that the two components are well-differentiated by their branch lengths (compare the ergodic average between $\sum int_1$ and $\sum int_2$, or between $\sum ext_1$ and $\sum ext_2$). The polymorphic class evolves under a tree with a total branch length around 10 times as long as the total length of the monomorphic class. Polymorphic sites accumulate a greater number of substitutions and so, the total evolutionary divergence, or total branch length, in the polymorphic class is much longer than that in the monomorphic class. The ergodic average of total interior length for the monomorphic component is 0.1056. Such a small value for the sum of six interior branch lengths suggests that the phylogenetic tree for the monomorphic class is nearly unresolved, as expected.

In this run, the acceptance rates were: 0.0046 for candidate phylogenetic trees; 0.5686 for branch-length updates; 0.5017 for rates of substitution; 0.3210 for stationary probabilities; and 0.2235 for allocation variables. The run took 11.8 hours to be completed.

parameter	*j=1*		*j=2*	
	ergodic average	$\hat{\tau}$	*ergodic average*	$\hat{\tau}$
$r_{(AC,j)}$	0.0318	121	0.0362	76
$r_{(AG,j)}$	0.3681	137	0.5347	260
$r_{(AT,j)}$	0.0618	52	0.0329	111
$r_{(CG,j)}$	0.0431	107	0.0622	244
$r_{(CT,j)}$	0.4751	123	0.2924	246
$r_{(GT,j)}$	0.0197	56	0.0414	238
$\pi_{(A,j)}$	0.2631	107	0.4045	191
$\pi_{(C,j)}$	0.2530	66	0.3388	137
$\pi_{(G,j)}$	0.1580	26	0.0534	109
$\pi_{(T,j)}$	0.3257	49	0.2032	173
ω_j	0.5529	39	0.4471	39
$\sum int_j$	0.1056	33	1.0851	32
$\sum ext_j$	0.2719	77	3.0991	57

Table 7.4: Ergodic average of model parameters and estimated integrated autocorrelation time from an analysis of the *primate mtDNA* alignment with a two-component $Q+t$ mixture model. Here j refers to the mixture component and the notation $\sum int_j$ and $\sum ext_j$ refers to the total length of interior and exterior branches for component j.

The MAP tree, with posterior mass of 0.9575, is shown in Figure 7-9. During the MCMC run, only five trees were visited, out of a total of $135\,135$ in the tree space. This highlights the strong tree-signal that the *primate mtDNA* data contain. This estimated topology matches the published trees in [126, 63] and [112].

7.5 Are there really two kinds of sites?

In order to test whether a split of the primate data into two classes is reasonable, we fitted the primate mtDNA alignment with a one-component $Q+t$ mixture model. We obtained parameter estimates under this model and used these estimates to produce five synthetic datasets, each of the same dimension as the original alignment. We then counted the number of monomorphic sites in each synthetic dataset and obtained an average of 28.55% monomorphic sites across the five datasets. This average number is significantly less than in the original alignment, where the proportion of monomorphic positions is 43.69%. This result provided some evidence that the primate mtDNA alignment cannot be adequately explained by only one component and that this dataset contains a second class of sites that require to be explained by a different component.

To provide further support to our findings, we analysed one of the synthetic alignments with $k=2$. The $Q+t$ mixture model detected only one component underlying these data.

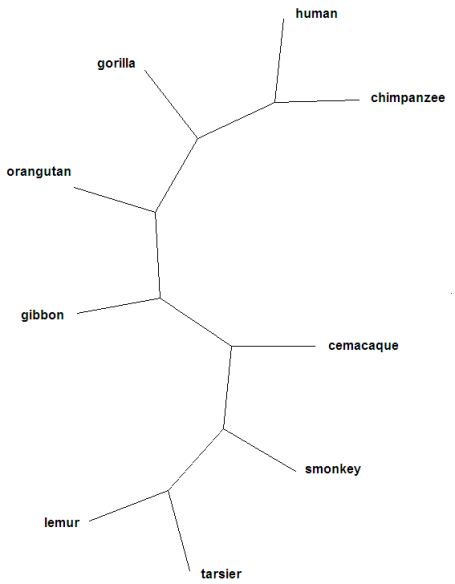


Figure 7-9: MAP tree for the *primate mtDNA* alignment. This topology has an estimated posterior mass of 0.9575 in the analysis in Section 7.4, and of 0.8887 in the analysis in Section 7.6. Here the lengths of the branches are meaningless as it is only the branching structure that is of interest. Crab-eating macaque is abbreviated as 'cemacaque', common squirrel monkey as 'smonkey', Philippine tarsier as 'tarsier' and ring-tailed lemur simply as 'lemur'. (See Section 4.5 for a criticism of the MAP tree.)

This result nicely illustrates what we discussed in Section 7.2.4. If we regarded this 'single detected' component as a polymorphic class (since most of the sites here are polymorphic), there are several monomorphic sites that nevertheless get allocated to this evolving class with high probability. However, when the proportion of monomorphic sites increases, as in the original primate mtDNA alignment, one component is not enough anymore to fit the data and a split into two evolutionary components is therefore required. A split of the primate mtDNA data into two classes –one mostly polymorphic and a second one mostly monomorphic– seems reasonable.

7.6 A three-component analysis

We analysed the primate mtDNA alignment with a $Q+t$ mixture model with three components. The tuning parameters and hyperparameter β were set to previously specified values (Table 7.3) and the length of the run was $40\,000$ iterations in total, with the initial $10\,000$ steps discarded as burn-in. The MCMC output exhibited a partition into three classes: one (mostly) polymorphic and two (mostly) monomorphic components. A close inspection of the histograms and traces of sampled values revealed that the two monomorphic components are not different. Figure 7-10 shows the histograms and traces of sampled total interior and exterior branch lengths. These plots exhibit a clear overlap between two of the three components; the two overlapping components correspond to the monomorphic classes. In Section 5.4, we discussed our criterion for deciding on a suitable number of mixture components. This criterion chooses the largest value of k such that all the mixture

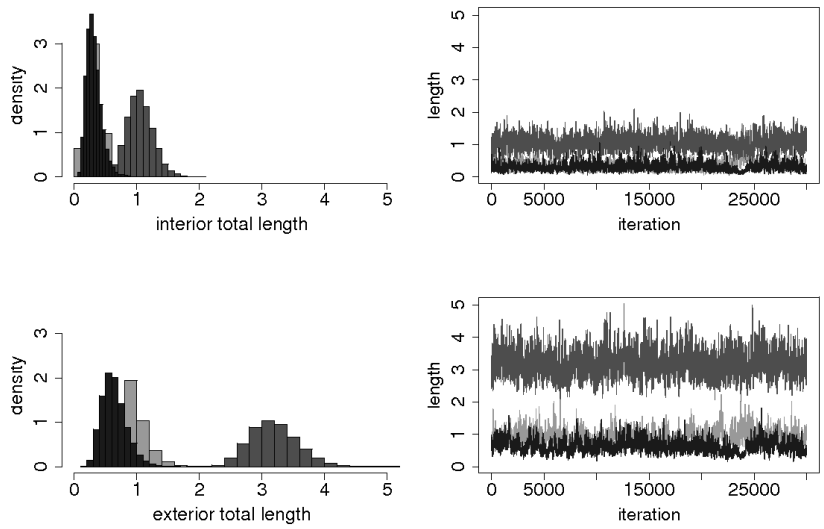


Figure 7-10: Histograms of the sampled total interior and exterior branch lengths for the analysis of the *primate mtDNA* alignment with a three-component mixture. The traces of sampled values are shown to the right of each histogram. The two components overlapping (in light grey and black) correspond to the monomorphic components.

components are different and all the associated mixture proportions are non-zero. Since there is no evidence to believe that the two monomorphic components are different, we decided on $k = 2$ as the adequate number of components to fit the primate mtDNA alignment.

We further verified our findings by running two more analyses on the primate mtDNA data with a three component mixture. In all cases we observed a three-class partition with two components overlapping. This provided additional support to the irrelevance of $k = 3$ in fitting this alignment.

7.7 Discussion

7.7.1 A different methodology

There are two keys ideas in our treatment of the *primate mtDNA* data which make it different from previously published methodologies. Firstly, no *a priori* knowledge of site membership to an evolutionary class is assumed. In contrast with Yang's [126] or Larget and Simon's [63] approaches, we do not rely on prior knowledge to classify the sites but rather infer their classification probability directly from the data. Secondly, we do not assume that a mixture with four components explains the alignment (as a $cp1$, $cp2$, $cp3$, and tRNA partition would suggest) but test for mixtures with as few as two or three components.

A mixture with two components explains the heterogeneity underlying this dataset in terms of the evolutionary differences between monomorphic and polymorphic sites. Sites which originate from the same type of position (either a monomorphic or a polymorphic

type) form clusters. As a by-product of this monomorphic/polymorphic partition, we can also differentiate the codon positions and the tRNA region. More variable codon positions, such as $cp1$ or $cp3$, have a large number of polymorphic sites and most of the sites originating from these positions cluster in one component. In contrast, more conserved regions, such as $cp2$ or tRNA, contain mostly monomorphic sites that cluster together in a common component.

A mixture with three components was found unnecessary to explain these data. Our choice of a two-component mixture as the most suitable model to fit the primate alignment disagrees with published results in which a four-class structure has been *a priori* assumed (e.g. [126, 63]). These findings further emphasise the relevance of the $Q+t$ mixture model in discovering structure underpinning phylogenetic data as part of a single inferential procedure. This, as opposed to imposing some class-structure which relies on *prior* knowledge about site membership to evolutionary classes.

We checked the MCMC output for label switching without finding any evidence of it. When appropriate, we used the commercially-available software $\mathrm{MrBayes}$ [51] to validate our estimates. Additionally, to get a feeling of the composition of the *primate mtDNA* alignment, we analysed it with the molecular evolutionary package MEGA [114]. This software produces frequency counts of each character type, i.e. number of As, Cs, Gs and Ts in the alignment. It also produces frequency counts of pairwise sequence changes (e.g. how many sites change from an A in sequence 1 to a C in sequence 2). These figures provide some clues of the true stationary probabilities $\boldsymbol{\pi}$ and substitution rates $\boldsymbol{r}$. In all cases, our estimates were supported by the evidence found by MEGA. Moreover, we checked that our estimate of the joint posterior distribution coincides with the chosen prior when no data is entered.

As a final remark, the genetic material of eukaryotes (the domain of like to which all animals belong) is known to very rarely undergo recombination of the type described in the footnote in Section 5.2.2 (which is commonly referred to as horizontal gene transfer; see, for example, [86]). Therefore, the assumption of a common topology among classes that the $Q+t$ mixture model makes is reasonable for this particular dataset.

7.7.2 An analysis of polymorphic sites

As a separate analysis, we investigated the presence of evolutionary heterogeneity in the polymorphic sites of the *primate mtDNA* alignment. To do so, we removed all the monomorphic sites from this alignment and fitted only the polymorphic positions with a two-component $Q+t$ mixture. Figure 7-11 shows the posterior classification probabilities to component 1, obtained from $30\,000$ iterations after burn-in of our MCMC sampler. In this figure, the $cp2$ region is the best differentiated. Nevertheless, sites in this alignment do not

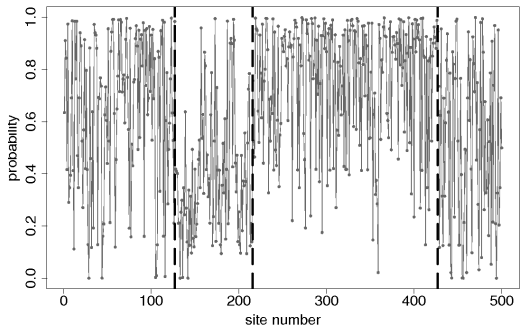


Figure 7-11: Estimated posterior classification probabilities to component 1 for polymorphic sites in the *primate mtDNA* alignment. Results obtained from an analysis with a two-component $Q + t$ mixture.

seem to be strongly heterogeneous; at least not in terms of the Q-matrix and the length of the branches.

An analysis of the polymorphic sites with a three-component mixture exhibits a nearly empty component during simulation which suggests, once again, that the common four-class assumption is inadequate.

Chapter 8

Evolutionary heterogeneity in *Borrelia burgdorferi*

8.1 Introduction

The purpose of this chapter is to test portions of the genetic material of the bacterium *Borrelia burgdorferi* for evolutionary heterogeneity. A phylogenetic mixture model that postulates distinct evolutionary classes can be used as a means to detect heterogeneity. If an analysis via the mixture model finds that more than one evolutionary class underlies the data, this would provide some evidence that the alignment contains heterogeneous phylogenetic signals.

Phylogenetic analysis is an essential tool in understanding the paths that govern bacterial evolution; issues involving host transmission, infection reservoirs and geographical origination of bacterial genetic variants (or *strains*) are more easily interpreted when examined in light of phylogenetic evidence. If the popular homogeneous model is used to derive phylogenetic conclusions, it is essential to verify that the data follow consistent evolutionary processes since an analysis of heterogeneous data with a homogeneous model could result in compromised inferences.

This chapter starts by presenting some background about the *B. burgdorferi* bacterium and stating the scientific question that motivated this study. It then introduces the datasets to be analysed and shows the results of these analyses, together with an overall discussion of our findings.

8.2 Background

B. burgdorferi is one of the bacterial species responsible for the most prevalent vector-borne disease in the temperate zone of the northern hemisphere, Lyme borreliosis (or Lyme disease) [71]. This condition is transmitted to humans by infected ticks when they attach to

the skin and feed on blood. The incidence in the United Kingdom has risen dramatically over the last few years, which might be due to a better awareness among the human population that, in turn, has led to a more efficient detection of the disease. However, there are other reasons that are also likely to explain the increment in reported human cases, such as colonisation of new habitats by wood-ticks, increasing tick population sizes and warmer year-round temperatures [71]. Most recent advances in our knowledge of Lyme borreliosis are related to a deeper understanding of the basic biology and ecology of its causative bacteria and the ticks that transmit it [110]. Biologists and other scientists around the world agree that, in order to fully understand the dynamics of the disease spread, it is crucial to understand the evolutionary paths of the bacterial strains that cause Lyme disease. This can be achieved by using phylogenetic tools.

Discrimination between different strains is the main purpose of *bacterial typing techniques*. Nowadays, typing methods are based on the practical application of molecular biology and they involve the direct analysis of bacterial DNA data (see [6, 74, 118] for an overview of typing techniques for microorganisms). Molecular typing methods should be highly discriminatory so that bacterial isolates assigned to the same strain are likely to descend from a recent common ancestor, while isolates that share a more distant common ancestor are not assigned to the same strain [70]. *Multilocus sequence typing* (MLST; [70]) achieves high levels of discrimination by selecting several genes that have been well conserved through evolution and analysing them in conjunction with one another. Nevertheless, the high discrimination that is gained by concatenating several genes might have an adverse effect if the same data are used for phylogenetic studies. A (homogeneous) phylogenetic analysis based on a concatenation of genes implicitly assumes that all genes share similar evolutionary properties. This is a strong assumption and, when untrue, inferences are a compromise between the different signals encoded in the data.

8.3 The scientific question

Our interest in *B. burgdorferi* comes from a publication by Margos et al. [71], in which they developed a novel MLST scheme for identifying *B. burgdorferi* strains based on eight housekeeping genes. Margos et al. referred to a number of other studies in which molecular data from different categories of loci[1], and not only from housekeeping genes, had been combined to identify *B. burgdorferi* strains (e.g. [8, 2]). If the different loci are incongruous in terms of their evolution, phylogenetic conclusions derived from a concatenation of conflicting *loci* will be distorted.

The purpose of our study is two-fold. Firstly, we aim to establish whether the eight housekeeping genes chosen by Margos et al. share evolutionary signals. If so, valid conclusions can be derived from *homogeneous* phylogenetic studies based on a concatenation of

[1] In molecular biology, a *locus* is a position on a chromosome, often used synonymously with *gene*.

those housekeeping genes. Biologists are interested in knowing this because it is the homogeneous phylogenetic model (with some possible variants) that is usually implemented in commercially-available software (e.g. MrBayes [51], BAMBE [108]). Secondly, we are interested in testing for heterogeneity between the housekeeping genes and some other categories of loci in use in the literature. Conflicting evolutionary signals between these genes would indicate inappropriateness of a phylogenetic study that analyses the genes concurrently under the homogeneous phylogenetic model. Instead, the model should account for the heterogeneity contained in the data by, for example, fitting a mixture of different evolutionary processes (e.g. the $Q + t$ mixture model).

The scientific question can be stated as follows: *Are the genes of interest congruent in evolution?* Only then valid phylogenetic inferences can be made under conventional methods that employ the homogeneous phylogenetic model to analyse a concatenation of several of these genes.

8.4 The *housekeeping gene* alignment

8.4.1 The data

The alignment assembled with eight chromosomal-located[2] housekeeping genes concatenated in the following order: $|clpA|clpX|nifS|pepX|pyrG|recG|rplB|uvrA|$ is referred to as the *housekeeping gene* alignment. This alignment contains the molecular data for the following *B. burgdorferi* strains: B31, IPT2, IPT19, IPT23, IPT39, IPT58, IPT69, IPT135, IPT137, IPT190, IPT191, IPT193, IPT198, NE49, Z41293 and Z41493.

In the original *housekeeping gene* alignment, 97% of the sites are monomorphic. Such a high level of monomorphicity dilutes the phylogenetic signal and produces star-like trees that provide ambiguous information about the evolutionary relatedness among taxa (Section 7.2). Given that polymorphic positions tend to follow less ambiguous patterns of evolution than their monomorphic counterparts, we removed all monomorphic sites and based our analysis only on polymorphic positions. (Indeed, there are a number of published phylogenetic studies of *B. burgdorferi* that also report including only polymorphic sites; e.g. [8, 2, 13].) There are, however, biases in parameter estimates introduced when removing monomorphic sites, as discussed in Section 7.2.4. In this section, however, we focus on discovering class-structure even when that results in biased parameter estimates. After thinning all monomorphic positions, the size of the *housekeeping gene* alignment is 16×121.

[2]The genetic material of bacteria is packaged into two types of structures: a *chromosome* which contains most of the DNA information and several independent extra pieces of DNA material termed *plasmids*. The genetic information contained in the chromosome is usually referred to as the *chromosomal DNA*; the extra DNA carried in the plasmids is called *plasmid-located DNA*. The total genetic material of *B. burgdorferi* is approximately 1.5×10^6 nucleotides long; of this, 60% is contained in the main chromosome and the rest is distributed among 21 plasmids [9].

The interest here is in testing whether the eight housekeeping genes share a consistent phylogenetic signal so that valid inferences can be drawn by fitting the homogeneous phylogenetic model, or similar variants. This set of housekeeping genes was used in the MLST scheme designed by Margos et al. [71]. This group of scientists were interested in validating the homogeneous model because it is the most commonly available in a number of commercially-available phylogenetic programs. For instance, $\mathrm{MrBayes}$ [51] is by far the most popular phylogenetic package for inference in a Bayesian setting, while PHYLIP [25], PAUP^* [113] and PAML [127, 129] are also widely used for maximum-likelihood inference. These programs implement the homogeneous phylogenetic model and some of them also implement the gamma model as a relaxation to homogeneity. In all cases, an analysis with these packages assumes that a single tree, set of branch lengths and Q-matrix are sufficient to characterise the evolutionary process across the entire DNA alignment (unless some prior knowledge about data partition is included). In order to obtain valid results under such a homogeneous formulation, it is necessary to demonstrate that all sites in the alignment are generated by the same evolutionary process. A natural way of doing so is to examine whether a mixture structure with more than one component is represented in the data or not. If the hypothesis of shared phylogenetic signal among the eight housekeeping genes is true, then only one mixture component should be detected and most of the sites will be classified to that single component.

8.4.2 A two-component analysis

We fitted the *housekeeping gene* alignment with a two-component $Q+t$ mixture model. The hyperparameter of the exponential prior distribution on a branch length was set to $\beta = 10$ (prior distribution (3.21)). Preliminary exploratory runs indicated that other choices of β yielded similar results. We estimated this model with the MCMC sampler described in Section 6.2, with tuning parameters set to the values given in Table 8.1. Quantity δ refers to the tuning value for the BLM proposal; σ refers to the standard deviation in the BLNA move; α_1 and α_2 are the tuning parameters of the ε-Dirichlet proposals for rates $\boldsymbol{r}$ and stationary probabilities $\boldsymbol{\pi}$, respectively; ε_θ is the size of the correction for the ε-Dirichlet proposal for $\boldsymbol{r}$ and $\boldsymbol{\pi}$; and ε_ω is the correction for the mechanism updating the mixture proportions. These values were chosen as the best performing ones, based on a number of initial exploratory runs.

Figures 8-1, 8-2, 8-3 and 8-4 summarise the results corresponding to $30\,000$ samples, following a burn-in period of $10\,000$ iterations. It is possible to observe satisfactory mixing patterns in all cases, with exception of the bottom plot in Figure 8-2. In that run, the chain encounters 'trapping states' and the mixing is slow. We will come back to discuss that specific case in the next section.

The posterior classification probabilities to the first component for sites $1, \ldots, 121$, i.e. $p(z_1 = 1|\boldsymbol{x})$, $p(z_2 = 1|\boldsymbol{x})$, $\ldots$, $p(z_{121} = 1|\boldsymbol{x})$, are shown in Figure 8-5. Most of the sites

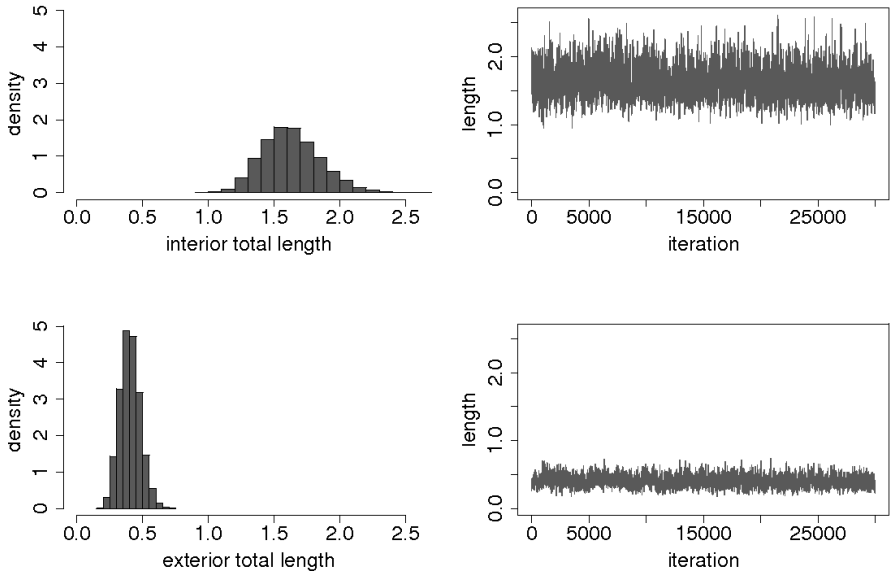


Figure 8-1: On the left-hand side, histograms of the posterior total length of interior and exterior branches for the *housekeeping gene* alignment fitted with a $Q + t$ mixture model, with $k = 2$. The traces of sampled values are shown to the right of each histogram.

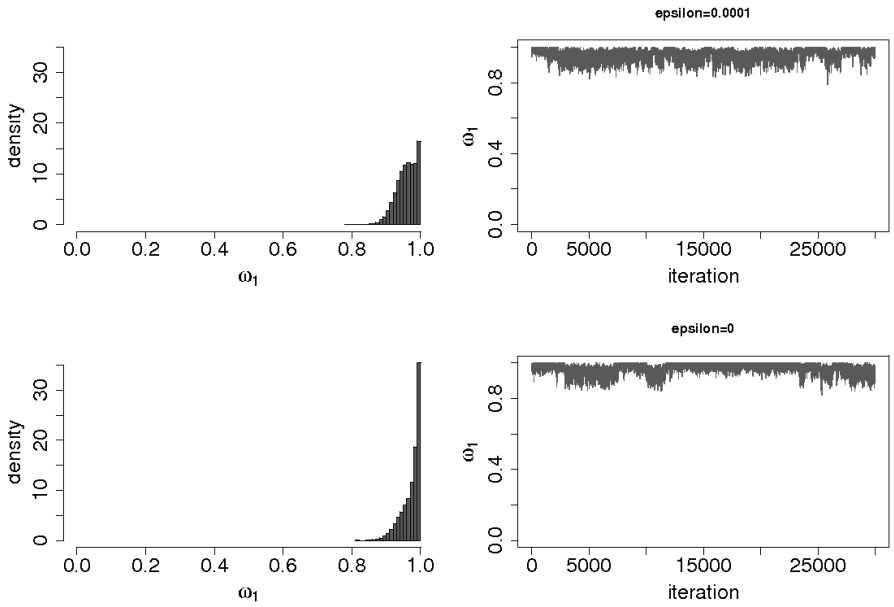


Figure 8-2: The top two plots correspond to an MCMC run for the *housekeeping gene* alignment, where the correction of the mechanism that updates the mixture proportions was set to $\varepsilon_\omega = 0.0001$. To the left, the histogram of the posterior mixture proportion ω_1 with a traceplot of the sampled values shown to its right. The bottom two plots correspond to an MCMC run with $\varepsilon_\omega = 0$. In this case, the traceplot exhibits evidence of 'zero-stickiness' of the complementary component.

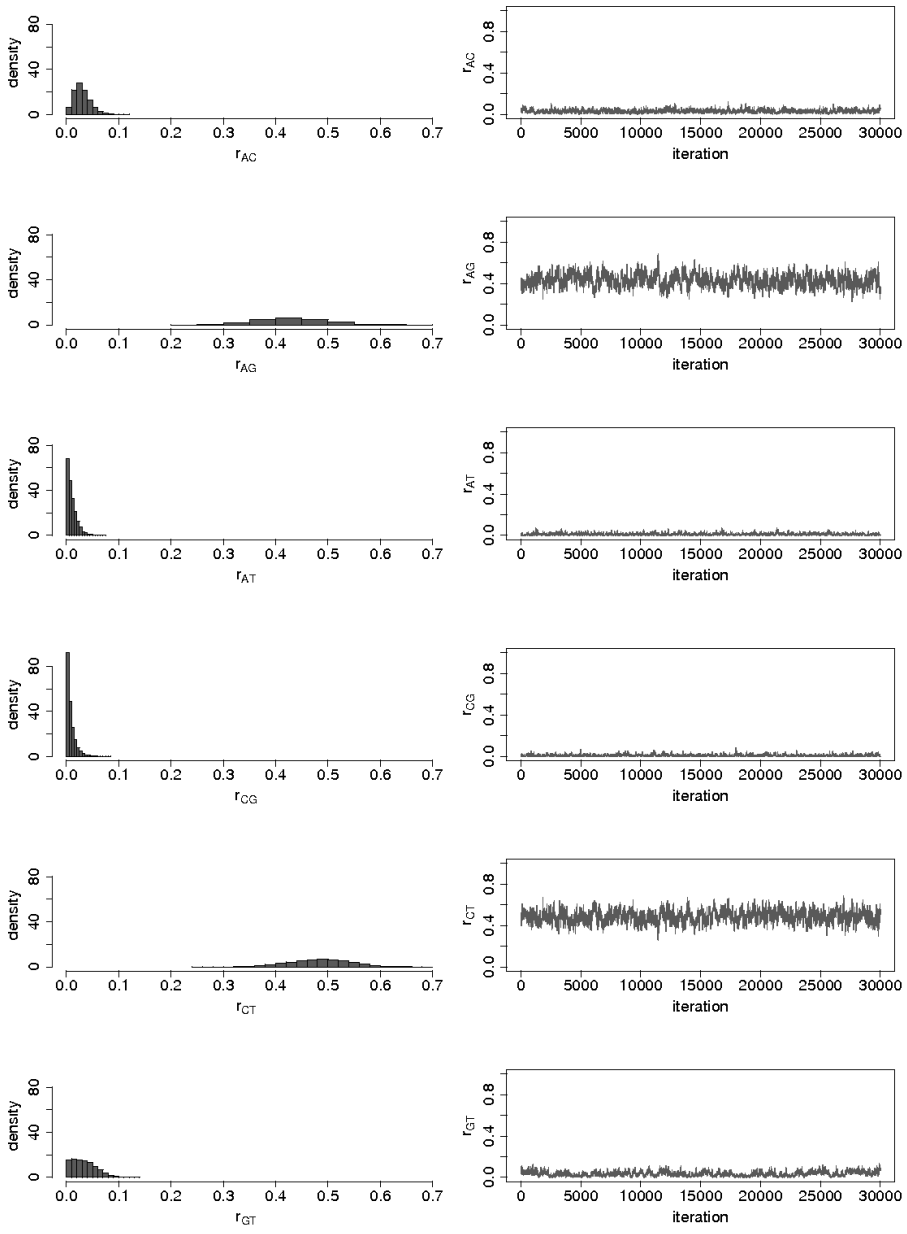


Figure 8-3: Histograms of posterior substitution rates for the *housekeeping gene* alignment fitted with a $Q + t$ mixture model, with $k = 2$. The traces of sampled values are shown to the right of each histogram.

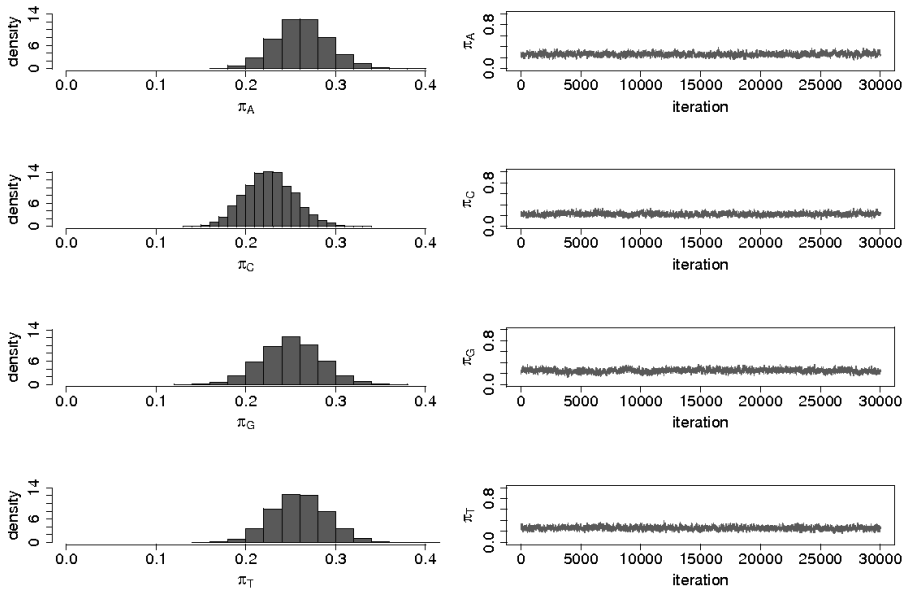


Figure 8-4: Histograms of posterior stationary probabilities for the *housekeeping gene* alignment fitted with a $Q+t$ mixture model, with $k = 2$. The traces of sampled values are shown to the right of each histogram.

are consistently allocated to the first component with high probability. This suggests that the eight housekeeping genes evolve under consistent rules and, therefore, a homogeneous model would appropriately fit the *housekeeping gene* alignment.

In Table 8.2, we report the ergodic average of model parameters and the estimated integrated autocorrelation times, for the first component. We can interpret the values from the column headed '$\varepsilon_\omega = 0.0001$' as follows. The estimates for the substitution rates $r_{(AG,1)}$ and $r_{(CT,1)}$ are noticeably higher than any other rate of substitution estimate, as expected (Section 3.3.3). Of these two rates, the one that changes $C \rightarrow T$ is the highest one. This bias has been observed before in the genomes of some bacteria and its cause is still a subject of investigation [99]. In this table, the estimated total length of interior

tuning parameter	*value*
δ	1.30
σ	0.04
α_1	800.00
α_2	600.00
ε_θ	0.0001
ε_ω	0.0001

Table 8.1: Tuning parameters for the update mechanisms used during the MCMC run for analysing the *B. burgdorferi* data. Parameter δ tunes the BLM move; σ tunes the BLNA move; α_1 and α_2 tune the εDP moves for substitution rates and stationary probabilities, respectively; ε_θ is the size of the correction for the εDP move and ε_ω is the correction for the mechanism updating the mixture proportions.

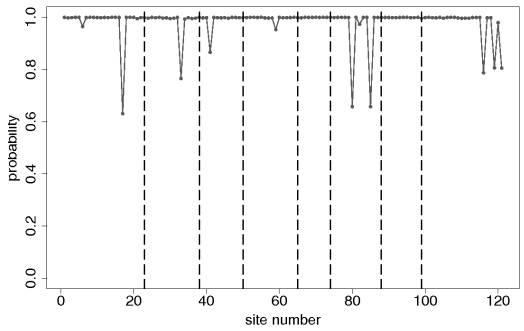


Figure 8-5: Posterior classification probabilities to the first component for the *housekeeping gene* alignment, i.e. $p(z_1 = 1|\boldsymbol{x})$, $p(z_2 = 1|\boldsymbol{x})$, ..., $p(z_{121} = 1|\boldsymbol{x})$. The alignment was fitted with a $Q + t$ mixture model with $k = 2$. The boundaries between the eight different genes have been marked as dotted lines.

branches, $\sum int_1$, is larger than that of exterior branches, $\sum ext_1$. Obtaining an ergodic average of 1.6321 for the sum of 13 interior branch lengths indicates that the tree is reasonably well resolved and far from being a 'star-like' tree (Section 7.2.2). These estimates are conditioned on all the sites in the *housekeeping gene* alignment being truly polymorphic. To recover the 'unconditioned' estimates we would have to scale these 'biased' estimates with a correction factor (Section 7.2.4). Nevertheless, the principle of site classification via the $Q + t$ mixture model is demonstrated and statements such as "$\hat{r}_{(AG,1)}$ and $\hat{r}_{(CT,1)}$ are noticeably higher than any other rate of substitution estimate" hold either for conditioned or unconditioned estimates.

The acceptance rates for the run were: 0.4772 for candidate phylogenetic trees; 0.7880 for branch-length updates; and 0.7142 and 0.7160 for rates and stationary probabilities, respectively. Candidate site allocations were accepted with rate 0.0189 (i.e. less than 2 candidate allocations accepted out of every 100 proposed). The low rate for allocation variables should not be a matter of too much concern if we interpret it in light of a strong signal contained in the data about site membership. It is clear for a site which component generated it so that it is unlikely to be allocated to a different component. The running time was 160 minutes.

Tree estimate

The MAP tree, with posterior mass of 0.0005, is shown in Figure 8-6(a). During the MCMC run, $12\,603$ trees were visited out of a total of 2.13×10^{14} in the tree space. The low posterior support of this tree highlights the diluted tree-signal in the *housekeeping gene* alignment. We do not know the biological reason for this, but it might be due to the similarity between bacterial strains.

For comparison, we have included in Figure 8-6(b) the *consensus tree* found by the program $\mathrm{MrBayes}$ [51]. Once having shown that the eight housekeeping genes share evo-

parameter	$\varepsilon_\omega = 0.0001$		$\varepsilon_\omega = 0$	
	ergodic average	$\hat{\tau}$	*ergodic average*	$\hat{\tau}$
$r_{(AC,1)}$	0.0298	67	0.0299	65
$r_{(AG,1)}$	0.4298	137	0.4211	400
$r_{(AT,1)}$	0.0105	48	0.0110	60
$r_{(CG,1)}$	0.0083	52	0.0076	51
$r_{(CT,1)}$	0.4870	106	0.4853	113
$r_{(GT,1)}$	0.0343	459	0.0447	967
$\pi_{(A,1)}$	0.2676	68	0.2623	49
$\pi_{(C,1)}$	0.2297	72	0.2264	113
$\pi_{(G,1)}$	0.2437	109	0.2513	283
$\pi_{(T,1)}$	0.2588	50	0.2598	39
ω_1	0.9588	**372**	0.9730	**1254**
$\sum int_1$	1.6321	26	1.6072	23
$\sum ext_1$	0.4040	49	0.4082	45

Table 8.2: The ergodic average of model parameters and the estimated integrated autocorrelation time, $\hat{\tau}$, for an analysis of the *housekeeping gene* alignment with an ε-corrected proposal for mixture proportions ($\varepsilon_\omega = 0.0001$) and for an analysis without a correction ($\varepsilon_\omega = 0$). When no correction is used in the proposal, the correlation between samples is very high, causing a large $\hat{\tau}$ for sampled values of ω_1 (highlighted in bold numbers). Adding a small correction $\varepsilon_\omega = 0.0001$ significantly improves mixing. The notation $\sum int_1$ and $\sum ext_1$ refers to the total length of interior and exterior branches for component one, respectively.

lutionary rules between them, it is valid to analyse this alignment with conventional phylogenetic software. Instead of picking a single 'most likely' tree from the MCMC output, MrBayes returns a consensus tree. Constructing a consensus tree is a way of choosing common elements among the stream of sampled trees during simulation. In Figure 8-6(b) for example, the branch that leads to the *subtree* $(IPT193, IPT198)$ has a posterior support 0.96. This means that 96% of the trees used for inference showed the same two strains, $IPT193$ and $IPT198$, dangling together from a common interior branch.

The tree in Figure 8-6(b) displays an unresolved (or star-like) subtree at the top. This suggests that the data do not contain enough information to decide which pairs of strains among $B31, IPT191, IPT190, IPT137$, $IPT135, IPT69, IPT23$ and $IPT2$ are closest neighbours. The two trees Figure 8-6 show reasonable correspondence.

8.4.3 Performance of proposals for mixture proportions

Section 6.2.1 presented a discussion about the inability of the proposal for mixture proportions, in its 'non-corrected' form (6.2), to stop the chain from falling into trapping states near the zero-boundary. When trapped, the chain might spend several iterations at

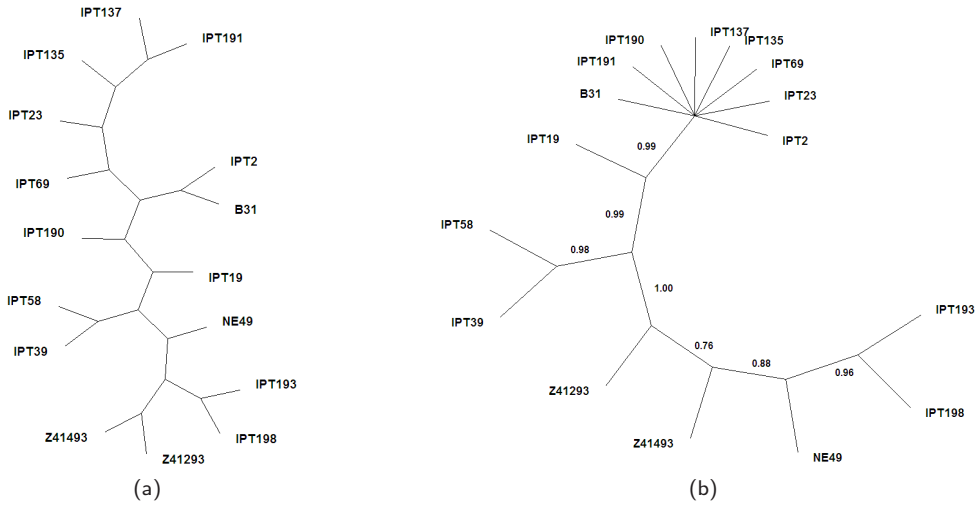


Figure 8-6: (a) MAP tree for the analysis of the *housekeeping gene* alignment with a two-component $Q + t$ mixture model. This tree has an estimated posterior mass of 0.0005. (See Section 4.5 for a criticism of the MAP tree.) (b) Consensus tree generated with the software package $\mathrm{MrBayes}$, with the posterior support for the interior branches indicated by the numbers. The lengths of the branches in both trees are meaningless as it is only the branching structure that is of interest.

the zero-neighbourhood before being able to escape. This phenomenon has detrimental effects on the mixing of the chain and the solution proposed, back in Section 6.2.1, consists of offsetting the centre of the Dirichlet distribution by a small quantity $\varepsilon_\omega > 0$. Here we investigate the performance of this ε-correction by comparing the previous run, where $\varepsilon_\omega = 0.0001$, with a run in which $\varepsilon_\omega = 0$.

Consider an MCMC sampler that estimates a two-component $Q+t$ mixture model when fitted to the *housekeeping gene* alignment. For this analysis, all the settings are identical to those discussed above, including starting points and length of the run (see Table 8.1). The only exception is that, this time, $\varepsilon_\omega = 0$. The traceplot at the bottom right of Figure 8-2 shows the sampled values for ω_1. This traceplot displays many instances in which the chain is stuck at values near one. This is the consequence of a nearly-empty component 2 which, in turn, causes that the variance of the Dirichlet distribution from which new values ω_2' are sampled, is nearly zero. The chain is not able to escape from the zero-boundary (in the case of ω_2) or the one-boundary (in the case of ω_1) and spends several iterations stuck in this region.

The column headed '$\varepsilon_\omega = 0$' in Table 8.2 reports the empirical averages and the estimated integrated autocorrelation times from this analysis. The different mixing behaviours between the cases $\varepsilon_\omega = 0.0001$ and $\varepsilon_\omega = 0$ can be verified in the estimated integrated autocorrelation time for ω_1. When no correction is used in the proposal, the correlation between samples is very high, causing a large $\hat{\tau}$. Adding an ε-correction to the proposal prevents the chain from falling into trapping states and improves its mixing, with the advantage of adding no extra computational cost (the running time for this analysis was also 160 minutes).

8.4.4 Consistency of results

As an additional exercise, we analysed the original *housekeeping gene* alignment (of size 16×4785 including both monomorphic and polymorphic positions) with a two-component $Q+t$ mixture model. The method allocated most of the sites to a single component in agreement with the results previously presented. In order to investigate why our classification method did not isolate monomorphic sites into one class and polymorphic sites into another for this particular set of data, we fitted the monomorphic/polymorphic *housekeeping gene* alignment with a one-component $Q+t$ mixture model. We obtained parameter estimates under this model and used these estimates to produce a synthetic dataset. We observed the same proportion of polymorphic positions in the synthetic alignment as in the original alignment (only 3%). This result suggests that, in this case, both monomorphic and polymorphic positions belong to a single evolutionary class and, even though this class is almost a 'non-evolving' one, some substitution events still occur by chance to give rise to (the very few) polymorphic positions. We further fitted the synthetic alignment with a two-component $Q+t$ mixture and, once again, discovered a single evolutionary class underpinning these data.

8.5 The *housekeeping gene*|*ospC* alignment

8.5.1 The data

We assembled a new alignment by combining the eight housekeeping genes from above and an extra gene, called the *ospC* gene. The latter is found in the plasmid of *B. burgdorferi* and is a gene in charge of coding for a protein called protein C. The *housekeeping gene*|*ospC* alignment, as we refer to it, is a concatenation of nine genes that physically looks like this: $|clpA|clpX|nifS|pepX|pyrG|recG|rplB|uvrA|ospC|$. Originally, the *ospC* gene has 51% of its sites monomorphic, which is not excessive. However, the high proportion of monomorphic sites in the housekeeping genes (97% of monomorphic sites) led us to remove all monomorphic sites from the *housekeeping gene*|*ospC* concatenation and analyse only the polymorphic positions. As previously discussed, there are biases in parameter estimates introduced when removing monomorphic sites and it is necessary to use a correction factor to obtain unbiased estimates (Section 7.2.4). In this section, as before, our main interest is discovering class-structure even when that may bring, as a consequence, biased phylogenetic estimates. After thinning monomorphic sites, the size of the final alignment is 16×353. The housekeeping region occupies the first 121 sites and sites $122-353$ correspond to the *ospC* gene.

The purpose here is to examine a mixture structure with $k\geq 2$ components, corresponding to the biological hypothesis of 'evolutionary heterogeneity between the chromosomal-located housekeeping genes and the plasmid-located *ospC* gene'. Plasmid-located genes often encode for traits that are advantageous but not essential to the bacterium. This

secondary role results in a greater chance to undergo changes through evolution, as compared to other more important, highly-conserved genes such as the housekeeping genes [9]. The well-defined and different biological roles between the housekeeping genes and the *ospC* gene should be adequately explained by a mixture model. Sites that originate from the same gene category (either a chromosomal or plasmid-located category) should share evolutionary rules and, therefore, constitute one evolutionary component.

8.5.2 A two-component analysis

We fitted the $Q+t$ mixture model with $k=2$ components to the *housekeeping gene*|*ospC* alignment, with the hyperparameter of the exponential prior distribution on a branch length set to $\beta = 10$ (prior distribution (3.21)). We estimated this model with the MCMC sampler described in Section 6.2, with tuning parameters set to the values shown in Table 8.1. We report results corresponding to $60\,000$ samples, following a burn-in period of $20\,000$ iterations. Both the values for the tuning parameters and the length of the run were chosen from several exploratory runs that helped monitoring convergence to stationarity and good mixing performance of the chain.

Figures 8-7, 8-8 and 8-9 summarise the results of the analysis. These plots exhibit clear evidence of a two-component structure of the data.

The posterior classification probabilities to the first component, i.e. $p(z_1 = 1|\boldsymbol{x})$, $p(z_2 = 1|\boldsymbol{x})$, $\ldots$, $p(z_{353} = 1|\boldsymbol{x})$, are shown in Figure 8-10. High probabilities of allocation to this component match with the *ospC* region, which suggests an association of component 1 with the *ospC* gene and of component 2 with the housekeeping genes. Therefore, in the following we refer to component 1 as the *ospC component* and component 2 as the *housekeeping-gene component*.

Once having identified the nature of the components, we return to Figures 8-7, 8-8 and 8-9 for an interpretation. In Figure 8-7, the component in light grey corresponds to the *ospC component* and the component in dark grey to the *housekeeping-gene component*. This may seem counterintuitive since a class accumulating a large number of nucleotide substitutions (such as the plasmid-located, hypervariable *ospC* gene) should have longer branch lengths than the housekeeping-gene class, which accumulates nucleotide changes relatively slowly. In other words, the component in light grey should be localised to the right of the dark component in the two histograms at the top of this figure. The fact that this does not occur can be explained by the removal of the monomorphic sites. But there is a second potential reason, namely, the common tree topology that the $Q+t$ mixture model imposes across the housekeeping-gene and the *ospC* regions when analysed as a concatenation *housekeeping*|*ospC*. An inspection of the MCMC output for the individual exterior branches from the analysis of the *housekeeping gene* data (Section 8.4.2) compared to the current analysis shows a good correspondence between all exterior branches except

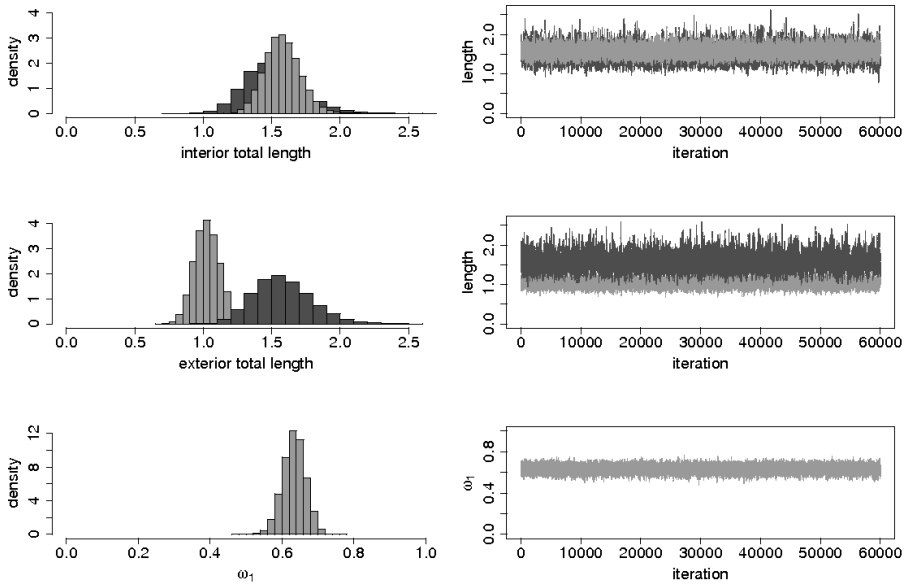


Figure 8-7: On the left-hand side, histograms of the posterior total length of interior branches, total length of exterior branches and mixture proportion ω_1 for the *housekeeping gene*|*ospC* alignment fitted with a $Q + t$ mixture model, with $k = 2$. The traces of corresponding sampled values for the parameter are shown to the right of each histogram. The component in light grey corresponds to the *ospC component* and the component in dark grey conforms to the *housekeeping-gene component*.

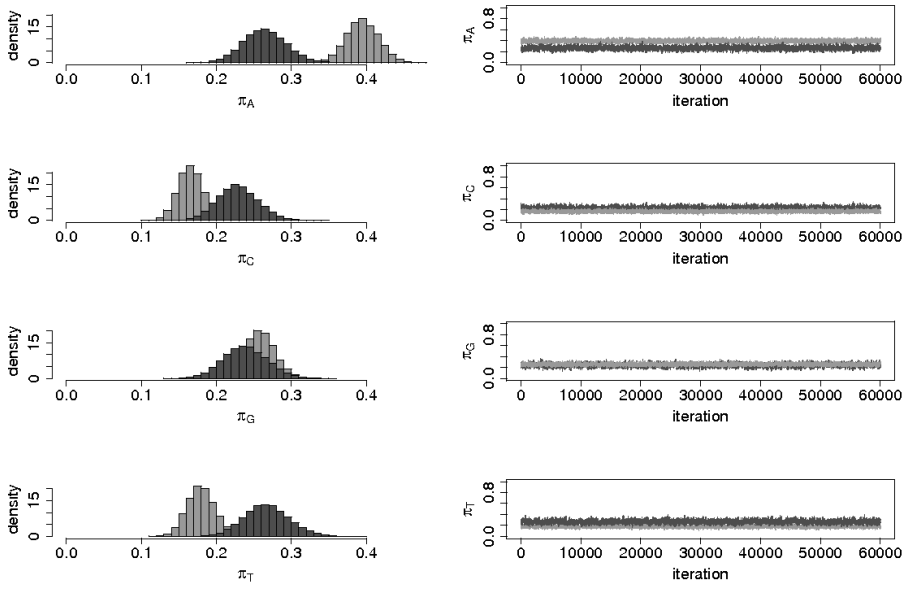


Figure 8-8: Histograms of posterior stationary probabilities for the *housekeeping gene*|*ospC* alignment fitted with a $Q + t$ mixture model, with $k = 2$. The traces of corresponding sampled values for the parameter are shown to the right of each histogram. The component in light grey corresponds to the *ospC component* and the component in dark grey conforms to the *housekeeping-gene component*.

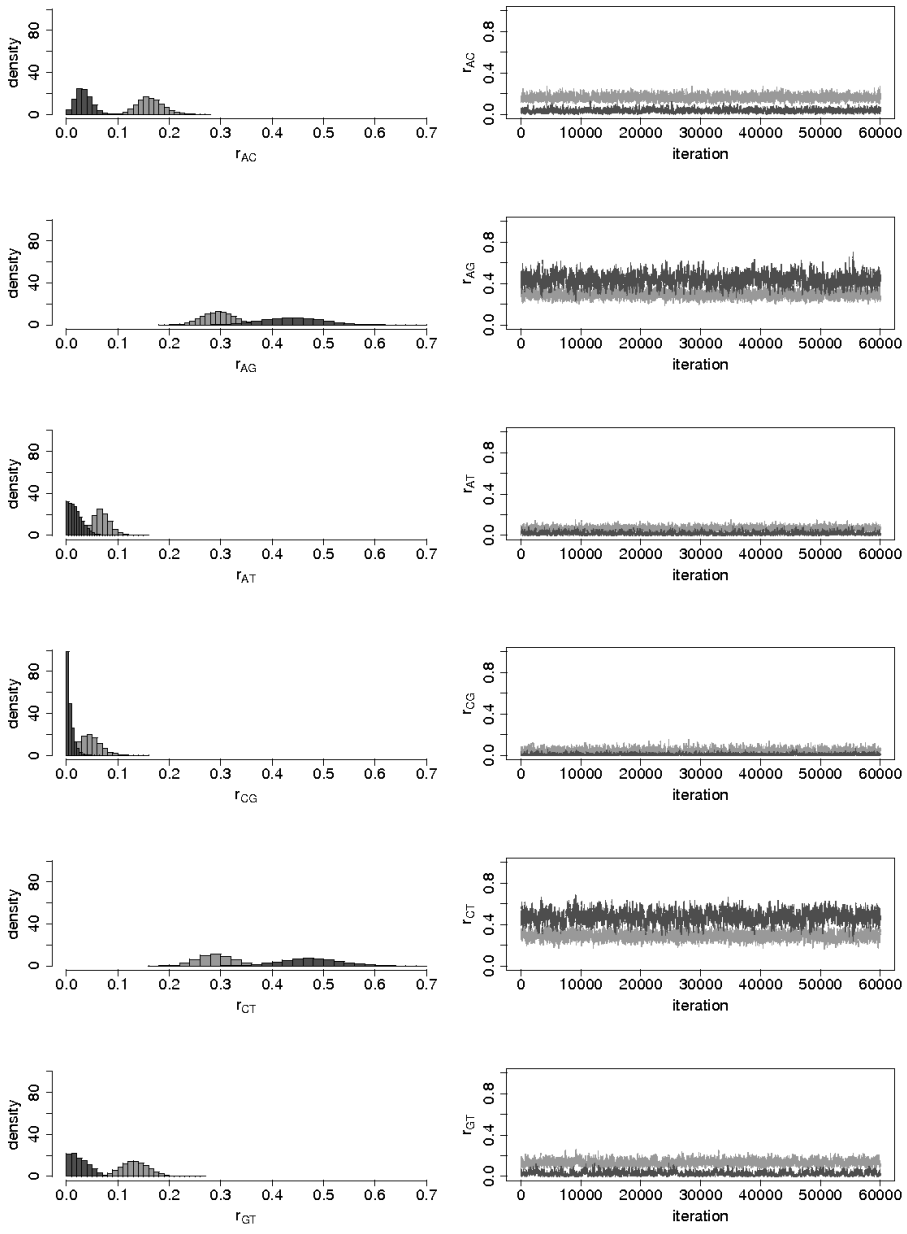


Figure 8-9: Histograms of posterior substitution rates for the *housekeeping gene*|*ospC* alignment fitted with a $Q+t$ mixture model, with $k=2$. The traces of corresponding sampled values for the parameter are shown to the right of each histogram. The component in light grey corresponds to the *ospC component* and the component in dark grey conforms to the *housekeeping-gene component*.

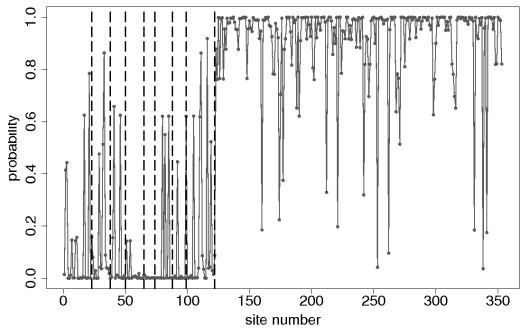


Figure 8-10: Posterior classification probabilities to the first component for the *housekeeping gene*|*ospC* alignment, i.e. $p(z_1 = 1|\boldsymbol{x})$, $p(z_2 = 1|\boldsymbol{x})$, ..., $p(z_{353} = 1|\boldsymbol{x})$. The data were fitted with a two-component $Q + t$ mixture model. The alignment was rearranged as $|clpA|clpX|nifS|pepX|pyrG|recG|rplB|uvrA|ospC|$. The boundaries between the eight different genes in the housekeeping region, and the *ospC* gene have been marked as dotted lines. Notice that high probabilities of allocation to this component match with the *ospC* region.

the branch leading to strain $Z41293$. In the analysis of the *housekeeping gene* data, the ergodic average of this branch is $\bar{t} = 0.0739$, while in the current analysis this average is $\bar{t} = 1.1822$ (although notice that the averages are based on different run lengths). The sampled lengths for this single exterior branch constitute the source of disagreement, but further research would be required to understand the reason for this in more detail.

An idea of the discrimination power of the two-component $Q + t$ mixture model when analysing a concatenation *housekeeping*|*ospC* can be obtained by comparing results from the analysis of the *housekeeping gene* alignment and those from the *housekeeping-gene component* in the current analysis. Notice, for example, the good correspondence between sampled rates of substitution in Figure 8-3 with the component in dark in Figure 8-9. A similar agreement can also be observed between Figure 8-4 and the component in dark in Figure 8-8.

Table 8.3 summarises the ergodic averages of model parameters and the estimated integrated autocorrelation times, $\hat{\tau}$. The first class ($j = 1$) conforms to the *ospC component* while class two ($j = 2$) to the *housekeeping-gene component*. The empirical averages for the rates of substitution agree with the bias that favours transitions (a substitution from $A \rightarrow G$ or $C \rightarrow T$) over transversions (any other possible substitution; Section 3.3.3). Also, the *ospC component* has an excess of A characters relative to the *housekeeping-gene component* (compare $\bar{\pi}_{(A,1)}$ with $\bar{\pi}_{(A,1)}$), which suggests A-richness in the *ospC* gene. The size of component 1, ω_1, adequately matches the proportion of *ospC* sites in the *housekeeping gene*|*ospC* alignment. There are 353 sites in total, of which 232 belong to the *ospC* region; this corresponds to 65.7% of the total sites. The ergodic averages of the interior total length in both components are large enough to correspond to reasonably well-resolved trees. As before, these estimates are conditioned on all the sites in the *housekeeping*|*ospC*

parameter	$j=1$		$j=2$	
	ergodic average	$\hat{\tau}$	*ergodic average*	$\hat{\tau}$
$r_{(AC,j)}$	0.1634	39	0.0340	61
$r_{(AG,j)}$	0.2968	44	0.4395	112
$r_{(AT,j)}$	0.0685	33	0.0187	84
$r_{(CG,j)}$	0.0465	71	0.0073	42
$r_{(CT,j)}$	0.2910	47	0.4732	107
$r_{(GT,j)}$	0.1336	81	0.0269	185
$\pi_{(A,j)}$	0.3953	24	0.2643	39
$\pi_{(C,j)}$	0.1663	28	0.2292	48
$\pi_{(G,j)}$	0.2586	30	0.2387	90
$\pi_{(T,j)}$	0.1797	28	0.2676	62
ω_j	0.6336	11	0.3664	11
$\sum int_j$	1.5785	14	1.5121	21
$\sum ext_j$	1.0296	17	1.5811	23

Table 8.3: The ergodic average of model parameters and the estimated integrated autocorrelation time, $\hat{\tau}$, for an analysis of the *housekeeping gene*|*ospC* alignment with a two-component $Q+t$ mixture. Here, class one ($j=1$) conforms to the *ospC component* while class two ($j=2$) to the *housekeeping-gene component*. The notation $\sum int_j$ and $\sum ext_j$ refers to the total length of interior and exterior branches for component j, respectively.

alignment being truly polymorphic but 'unconditioned' estimates may be calculated by applying a correction factor (Section 7.2.4).

The acceptance rates for this analysis were: 0.1748 for proposed phylogenetic trees; 0.6537 for branch-length updates; 0.5567 and 0.4640 for substitution rates and stationary probabilities, respectively, and 0.1121 for candidate site allocations. The running time was 800 minutes.

Tree estimate

The MAP tree, with posterior mass of 0.0341, is shown in Figure 8-11(a). During the MCMC run, 243 trees were visited out of a total of 2.13×10^{14} in the tree space. In the analysis of the *housekeeping gene* alignment with a two-component mixture, in Section 8.4.2, we found that the housekeeping data supports several competing trees. That is, the posterior distribution for trees in that case is fairly flat. This suggests that the MAP tree obtained from the *housekeeping gene*|*ospC* alignment mainly comes from the signal contained in the *ospC* region.

The low number of trees visited should not be a matter of concern since, despite the vast size of the tree space, the data support a relatively small set of trees. We verified this

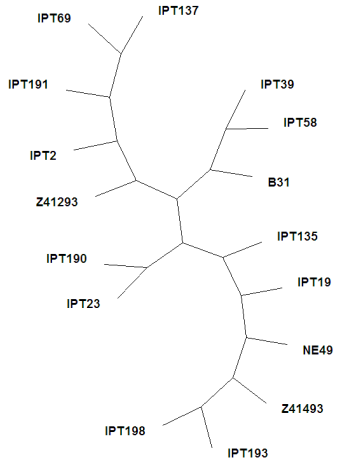


Figure 8-11: MAP tree for the analysis of the *housekeeping gene*|*ospC* alignment with a two-component $Q + t$ mixture model. This tree has an estimated posterior mass of 0.0341. (See Section 4.5 for a criticism of the MAP tree.)

by performing several runs starting from different trees and our sampler always found similar sets of likely trees. It is important to point out, however, that there were always several competing trees almost equally supported by the data. This is reasonable since we would not expect the *B. burgdorferi* strains to be highly differentiated so that only one tree described their evolutionary relations. Several similar trees seem to represent the phylogeny of these strains almost equally well. We also analysed the *housekeeping gene*|*ospC* alignment with MrBayes (under the homogeneous model) and found reasonable congruence in the set of supported trees.

8.5.3 A 'compromised' analysis

Figures 8-12, 8-13 and 8-14 summarise the results corresponding to $60\,000$ iterations, after $20\,000$ burn-in steps, of our MCMC sampler. The model fitted to the *housekeeping gene*|*ospC* data in this analysis is the $Q + t$ mixture with *one* component (which actually corresponds to the homogeneous phylogenetic model). The tuning parameters and other model specifications remained as in previous analyses (Table 8.1).

The purpose here is to show the 'compromised' inferences that a homogeneous model produces when fitted to data that arise from heterogeneous processes. Compare the plots below with the corresponding ones in Figures 8-7, 8-8 and 8-9. The sampled values under the homogeneous model in this section are a trade-off between the two evolutionary classes captured by an analysis with a two-component $Q + t$ mixture.

8.5.4 A three-component analysis

We fitted the *housekeeping gene*|*ospC* alignment with a $Q + t$ mixture model with $k = 3$ components. The tuning parameters and other model specifications remained as in previ-

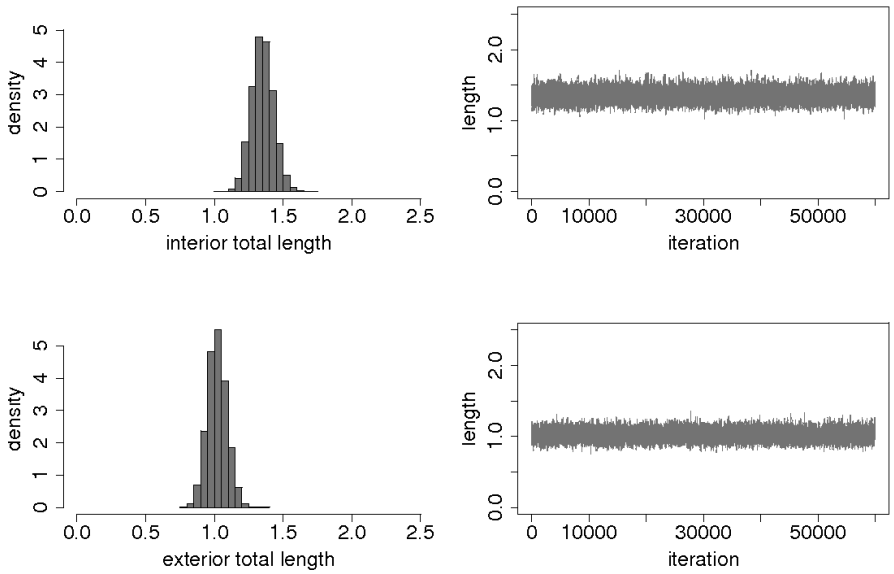


Figure 8-12: Histograms of the posterior total length of interior and exterior branches for the *housekeeping gene*|*ospC* alignment, fitted with a one-component $Q+t$ mixture model. The traces of corresponding sampled values for the parameter are shown to the right of each histogram. (Compare with the two-component analysis in Figure 8-7.)

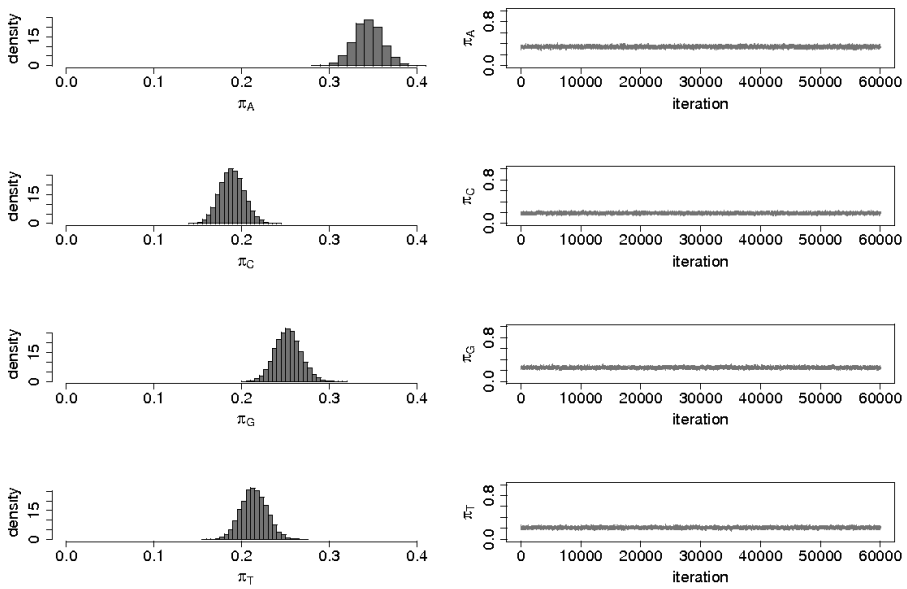


Figure 8-13: Histograms of posterior stationary probabilities for the *housekeeping gene*|*ospC* alignment, fitted with a one-component $Q+t$ mixture model. The traces of corresponding sampled values for the parameter are shown to the right of each histogram. (Compare with the two-component analysis in Figure 8-8.)

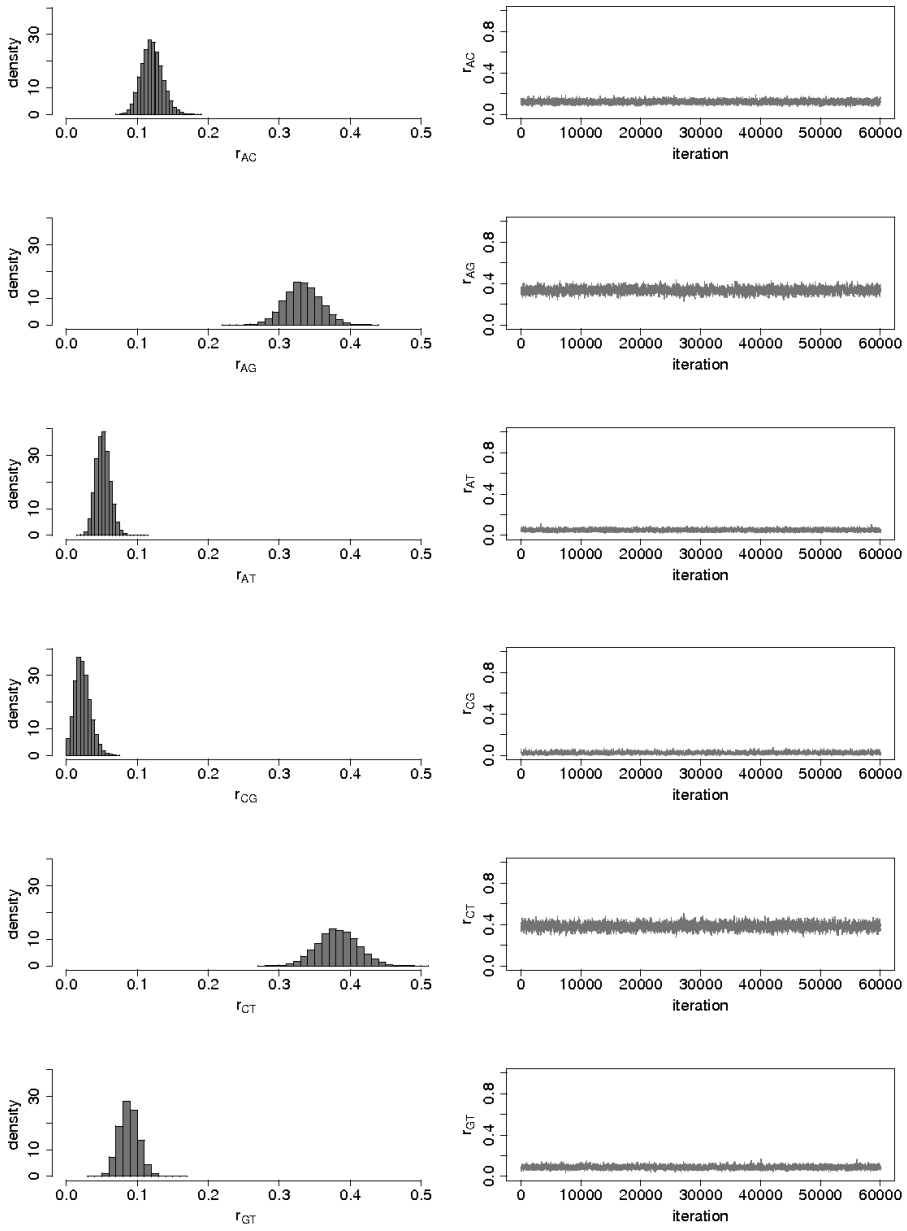


Figure 8-14: Histograms of posterior substitution rates for the *housekeeping gene*|*ospC* alignment, fitted with a one-component $Q + t$ mixture model. The traces of corresponding sampled values for the parameter are shown to the right of each histogram. (Compare with the two-component analysis in Figure 8-9.)

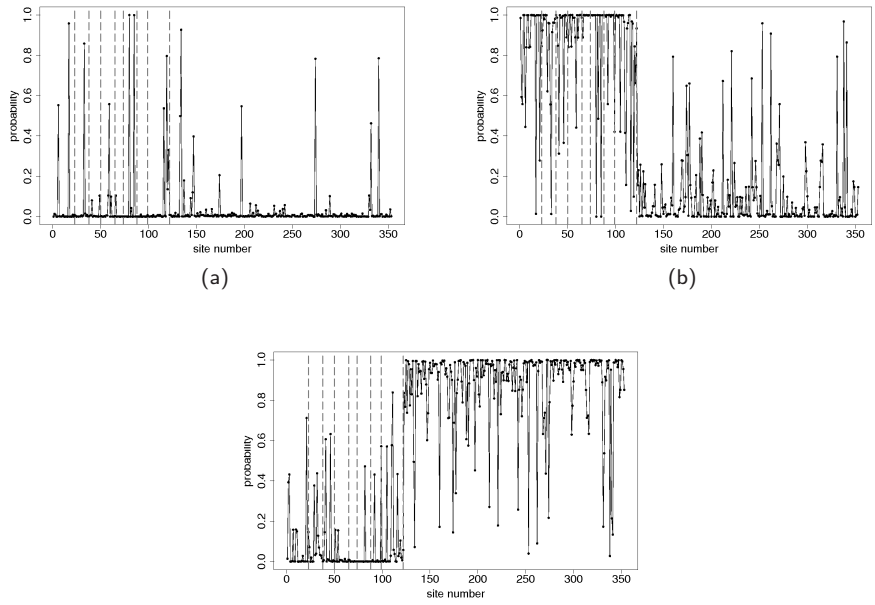


Figure 8-15: Posterior classification probabilities for the *housekeeping*|*ospC* alignment, with $k = 3$. (a) Component 1 exhibits very few sites with non-negligible probability of allocation to it, which suggests that this component is nearly empty. (b) Sites with high classification probability to component 2 are mostly located in the housekeeping region. (c) Sites with high classification probability to component 3 mainly originate from the *ospC* region. The boundaries between the eight different genes in the housekeeping region, and the *ospC* gene have been marked as dotted lines.

ous analyses (Table 8.1). We report results corresponding to $60\,000$ samples, following a burn-in period of $20\,000$ iterations.

The estimated posterior classification probabilities to the three components are displayed in Figure 8-15. The low probability of allocation to component 1 for most sites across the alignment, in Figure 8-15(a), suggests that this component is nearly-empty. The high probabilities of allocation in Figure 8-15(b) match with sites located within the housekeeping region. Thus, component 2 may be regarded as a *housekeeping-gene component*. Figure 8-15(c) displays high probabilities for sites in the *ospC* region, and thus component 3 may be associated with the *ospC* gene. These results reveal, once again, the ability of the $Q + t$ mixture to discriminate between the two evolutionary classes; one conforming to data arising from housekeeping genes and a second one describing *ospC*-gene observations.

The acceptance rates for this chain were: 0.1879 for proposed trees; 0.7246 for branch-length updates; 0.6238 and 0.5711 for substitution rates and stationary probabilities, respectively; and 0.0729 for candidate site allocations. The increase in the number of mixture components caused this analysis to be more computationally expensive than the previous ones. The running time was $1\,300$ minutes.

The nearly-empty component in an analysis with three components suggests that only two components underlie the *housekeeping*|*ospC* alignment. As a further test for the number of components, we fitted only the *ospC* region with a two-component $Q+t$ mixture. The results for this analysis (not shown) contain a mixture component that remains nearly empty throughout simulation, indicating that the *ospC* region can be adequately fitted with a 'mixture' with one component. In Section 8.4, we showed that the *housekeeping gene* region is best modelled by a one-component 'mixture'. Therefore, an alignment that concatenates the *housekeeping gene* and the *ospC* regions will be best fitted by a $Q + t$ mixture with two components.

8.6 Discussion

We shall begin by discussing the removal of monomorphic sites from the alignments. Thinning monomorphic positions is a measure we took for the Borrelia data in particular, while there are other applications for which removing monomorphic sites might not be required. Most of the problems we found arose when analysing the original *housekeeping*|*ospC* alignment (i.e. the alignment that includes both monomorphic and polymorphic sites). An analysis of this dataset with a two-component $Q + t$ mixture would classify the sites according to their monomorphic/polymorphic nature. Although a reasonable partition, we were interested in discovering class-structure beyond the obvious differences between monomorphic and polymorphic sites. To do so, we reasoned that keeping only polymorphic positions and testing for evolutionary heterogeneity among these 'evolving' sites could reveal further unseen structures. The objective of our analyses in this chapter has been to detect evolutionary differences among polymorphic sites at the cost of obtaining biased phylogenetic estimates (Section 7.2.4). Removing monomorphic positions was the approach we took but there is plenty of scope for improvement. For instance, one could specify instead stronger priors that restricted monomorphic sites to cluster into one class so that an analysis with a three-component mixture focused on discovering further partitions within the polymorphic class and did not attempt uninteresting partitions of the monomorphic component. This is something that has not been pursued as part of this thesis and could be considered as a topic for future research.

Checking the MCMC output for label switching before making any inference on component specific parameters, is important. We have examined the output without finding any evidence of unambiguous labelling of the components. This has allowed us to make valid inferences without the need to resolve any label switching.

The plots of posterior classification probabilities, in Figures 8-5, 8-10 and 8-15, all show some sites that have been classified to a different component than expected. An inspection of the nucleotide composition in the 'misclassified' sites suggests that this may be the cause. For instance, when discussing the results in Table 8.3, we highlighted the A-richness of the

ospC component relative to the *housekeeping-gene component* (compare $\pi_{(A,1)}$ with $\pi_{(A,2)}$ in that table). This could cause a site that belongs to the housekeeping region but that, for some reason, has a high proportion of A characters to be allocated to the *ospC component*.

Phylogenetic inference is a complex procedure and we consider that the level of discrimination between classes that the model achieves is good. The model can detect, for example where the *housekeeping gene* alignment ends and the *ospC* gene starts, as well as correctly classify most of the sites that belong to one or another category (Figure 8-10). The 'misclassification' of some sites could be taken as the cost of dealing with such complex models.

Several independent MCMC runs starting at different points showed convergence of the chain to the same distribution. There are different points of view in the literature, but some recommendations maintain that it is worthwhile comparing independent chains for similarity in distribution [38]. If the chains show differences, this means that stationarity has not been reached yet and that the simulation needs to be run for longer.

To get a feeling of the composition of each of the genes analysed in this chapter, we studied them first with the package MEGA [114]. This software produces frequency counts of each character type in the DNA alignment, i.e. number of As, Cs, Gs and Ts in the alignment. It also produces frequency counts of pairwise sequence changes (e.g. how many sites in sequence 1 change from A to C in sequence 2). These figures provide an idea of the true stationary probabilities $\boldsymbol{\pi}$ and substitution rates $\boldsymbol{r}$. In all cases, our estimates were supported by the evidence found by MEGA.

We also performed some runs in $\mathrm{MrBayes}$ [51]. When using this package, we analysed data known to be homogeneous (e.g. the *housekeeping gene* alignment or the *ospC* gene individually). This way we could compare some of our parameter estimates with those obtained in $\mathrm{MrBayes}$. In all cases we found a good correspondence.

The results from the analyses presented in this chapter have had direct relevance to biologists studying the dynamics of the spread of Lyme disease. In order to identify and monitor the origins, directions and diversity of the bacterial species that cause the disease, biologists make use of phylogenetic insights. The group of Margos and colleagues, working at the University of Bath, was interested in validating their published phylogenetic conclusions in [71]. These results were derived from the analysis of a concatenation of the eight housekeeping genes presented in this chapter, fitted with the homogeneous model. Valid phylogenetic inferences in this case rely on there being evolutionary consistency between the eight concatenated housekeeping genes. The results in this chapter have demonstrated evolutionary congruence between these genes and, therefore, provided statistical support to the work in [71].

Similar analyses to the ones described in this chapter could be routinely performed on concatenated data to discover structures among sites. If different classes are detected by the $Q + t$ mixture model, an interpretation of the underlying heterogeneity may be attempted by looking at the estimated posterior classification probabilities. The $Q + t$ mixture model is a powerful tool to reveal unknown class structures which, in turn, could lead to a deeper understanding of the processes shaping evolution.

Several results from this chapter have been presented at both national and international meetings (*Learning in Computational Systems Biology*, Imperial College, London, UK, April 2009; *X International Jena Symposium on Tick-borne Diseases*, Weimar, Germany, March 2009).

Chapter 9

Conclusions and further work

This thesis proposed a novel classification method for phylogenetic data. Our method assumes that a DNA alignment is characterised by multiple evolutionary classes, each class with a distinctive set of branch lengths and Q-matrix that generates the Markov process of nucleotide substitution. The conventional homogeneous phylogenetic model, on the other hand, postulates that all sites in the DNA alignment evolve under the same phylogenetic tree, set of branch lengths and Markov process of nucleotide substitution. It is well-known, however, that DNA evolves in a heterogeneous way, causing different sites in an alignment to obey different processes of evolution. The modelling of heterogeneous phylogenetic data has long been a subject of interest. A number of overall-rate models, hidden Markov models, change-point models and, more recently, mixture models, have been proposed in the literature as relaxations to the over-simplistic homogeneous model. Until now and as far as we are aware, none of these previously-proposed treatments have allowed for both the possibility of including a reasonable number of sequences in the analysis and a natural framework for classification. In contrast, the number of sequences that can be analysed by the $Q + t$ mixture model described in this thesis is limited only by the computational power available and the missing-data reformulation of the mixture enables classification of the sites as part of a single inferential procedure. Reformulating a finite mixture model as a problem of missing-data is a well-known strategy in the statistical literature when one of the objectives of the analysis is the classification of a set of observations. In this reformulation, the purpose is to infer the identity or label of the mixture component from which each site is drawn. In a phylogenetic context, each mixture component represents one evolutionary class. Once the identity of the component that generates each site is recovered, it provides us with a natural framework for classifying the sites.

The research underlying this thesis had three main aims. The first of these concentrated on providing a classification scheme for phylogenetic data based on a Bayesian finite mixture model. The second aim related to contributing MCMC methodology for the efficient estimation of the parameters of a phylogenetic mixture model. The third was concerned with the assessment of the performance of the classification scheme and the estimation

procedures when applied to real data. Conclusions on the results of each of these aims are discussed in detail below, beginning with the classification of sites via a Bayesian mixture model.

9.1 Site classification via Bayesian mixture modelling

The phylogenetic model that we propose for site classification postulates a mixture of different evolutionary components operating in nature to generate the phylogenetic data. Each component is characterised by a specific set of branch lengths and Q-matrix, while the phylogenetic tree is assumed common to all components. The length of a branch represents the *expected number of nucleotide substitutions* between the two connected nodes. It is a measure of the *amount* of evolutionary divergence between the taxa. On the other hand, the Q-matrix contains the *rules* that specify the Markov process of nucleotide substitution. Different rules generate different substitution processes, and different sets of branch lengths conform to processes that accumulate different amount of evolutionary divergence. Therefore, postulating a mixture model that includes multiple sets of branch lengths and multiple Q-matrices is a natural way of allowing for both heterogeneity in the *amount* and the *rules* of evolution.

Our approach characterises the evolutionary components according to sets of branch lengths and Q-matrices, but several other possibilities exist. One could instead specify a mixture on trees and branch lengths with a common-to-all-component Q-matrix (e.g. [121]), or a mixture on all parameters: tree, branch lengths and Q-matrix. Mixtures that include multiple trees may be computationally expensive and measures to overcome this may be needed. One approach would be to introduce a mechanism that constrains the tree space to only a few trees supported by the data. Applications where most of the trees have very little posterior support and can be effectively removed from the inference would be greatly benefited by an approach like this. A key reference to the idea of 'stochastic selection of supported trees' is Webb, Hancock and Holmes [121].

A restrictive but interesting alternative would be to construct a *mixture on overall-rates of substitution*. That is, a mixture that includes multiple overall-rates, one per each component, and a single tree, set of branch lengths and Q-matrix that are common to all components. The overall-rate for component j would act as a scaling factor of the set of branch lengths. Thus, different components would have trees with differently scaled branch lengths; one component might have a tree where all branch lengths have shrunk in the same proportion, while a different component could have a tree with 'equally enlarged' branches. Such a formulation would be more restrictive that the $Q + t$ mixture model described in this thesis, and would suffer from similar limitations to the homogeneous model. Nevertheless, it would allow for direct comparison with the popular *discrete-gamma model* [125] (Section 5.2.1). In a discrete-gamma formulation, the overall-rates are assumed to conform

to a Gamma distribution and all components are constrained to have equal relative size (expression (5.2)). In contrast, a mixture on overall-rates would allow more flexibility by letting the data support different relative sizes of the components. It would also let the overall-rates take any (positive real) value and not restrict them to conform to a Gamma distribution. A mixture on overall-rates would, nevertheless, incorporate as many extra parameters in the model as components in the mixture (relative to the homogeneous phylogenetic model), while a discrete-gamma model only adds one extra parameter. However, a simple reformulation of a mixture of this type as a missing-data problem would enable a natural framework for classifying the sites. Ultimately, the realism gained by a more parameter-rich but also flexible mixture model could outperform the simplicity and good features of the discrete-gamma model. At present we are not aware of any study that has systematically compared the discrete-gamma model with a mixture on overall-rates and we believe that this would be an interesting area of future research.

One potential criticism of our $Q + t$ mixture model is the assumption of a process of nucleotide substitutions ruled by the GTR matrix. In fact, there might be applications in which some GTR parameters are superfluous and lose their interpretation. For instance, if one of the mixture components has a negligible proportion of A characters allocated to it, performing inference on rates r_{AC}, r_{AG} and r_{AT} for that component will be meaningless. We have based our model on the GTR matrix but there is nothing in the nature of the model that restricts it to assume this matrix. The $Q+t$ mixture model can be readily extended to assume other models of nucleotide substitution that may fit the data better, i.e. to assume a HKY85 model instead of a GTR, for example (Section 3.3.4). Furthermore, it is possible to implement a reversible-jump MCMC [42] to move between different parametrisations of the model of nucleotide substitutions. So, if the current Q-matrix has a GTR parametrisation $\boldsymbol{\theta} = ((r_{AC}, r_{AG}, \ldots, r_{GT}), (\pi_A, \ldots, \pi_T))$, the MCMC sampler could propose at the next iteration a HKY85 parametrisation of the form $\boldsymbol{\theta}' = ((r_T, r_V), (\pi_A, \ldots, \pi_T))$, where $r_T = r_{AG} = r_{CT}$ denotes the rate of transitions and $r_V = r_{AC} = r_{AT} = r_{CG} = r_{GT}$ denotes the rate of transversions. Such a treatment would allow for mixtures with alternatively parametrised components. Some components may be indexed by parameter-rich Q-matrices, while others by simpler Q-parametrisations. Moreover, the adequacy of a model would no longer need to be tested by Bayes factors or other methods that tend to be cumbersome within a phylogenetics context, since the sampler itself would find the most well-supported model under the observed data. A key reference to phylogenetic model selection via reversible-jump MCMC is Huelsenbeck, Larget and Alfaro [50].

Another possible extension of our work is the fully Bayesian treatment of the $Q + t$ mixture model. That is, to model the number of components and the mixture component parameters jointly and to infer both. With such a treatment, we would no longer need to consider models for different numbers of components separately, nor use any post-simulation criterion to estimate the number of mixture components (Section 5.4). An MCMC approach

to estimation under this scenario would require the sampler to move between mixtures with a varying number of components and hence reversible-jump MCMC [42] could be employed. The MCMC sampler in this case could make use of moves that split one mixture component into two or combine two into one (*split/combine* moves), and moves that 'give birth to' or remove an empty component (*birth/death* moves). A useful reference in this topic is Richardson and Green [94].

The classification method that we have presented not only accounts for the heterogeneity underlying the data but also allows us to identify and interpret the classes represented in the data. This method is a step forwards from previously-published methodologies in which the formulation would account for heterogeneous data but would not enable site classification, at least not directly. Furthermore, since the basic component of our model is the well-known homogeneous formulation, we designed our estimation methods based on proposed MCMC algorithms in the phylogenetics literature but with some improvements where pertinent. The next section presents conclusions regarding the MCMC algorithms we investigated and devised as part of this research.

9.2 MCMC methodology

Chapter 4 investigated existing MCMC algorithms for estimating the homogeneous phylogenetic model. One of the most popular proposal mechanisms for updating tree and branch lengths is the so called LOCAL proposal by Larget and Simon [63]. This mechanism proposes new branch lengths in a neighbourhood of the tree which may or may not generate a new tree as a by-product. Such a simultaneous, or *en bloc*, update of tree and branch lengths is shown to cause bad mixing of the chain in applications where only a few trees are supported by the data. This is because most candidate trees will keep being legitimately rejected at most iterations at the same time that reasonable candidate branch lengths are also (unfairly) rejected. Section 4.4 proposed an alternative mechanism that performs separate updates of tree and branch lengths, and showed how this significantly improves the estimation performance of the chain.

Section 4.4.4 demonstrated the benefits of employing two different types of moves for updating branch lengths, BLM and BLNA. The analysis of a synthetic dataset suggested good estimation performance of BLNA at regions of the state space near the zero-boundary and good estimation performance of BLM for longer branch lengths. The derivation of the variance of a candidate length generated via BLM showed that the step-size of BLM depends on the value of the current branch length whereas the step-size of BLNA does not. We believe that this could be related to the different performances of the proposals at different regions of the state space. However, these beliefs are only based on results obtained from one dataset and further research would be required to determine whether these conclusions hold more generally. In our MCMC sampler, we chose to alternate between

BLM and BLNA at even and odd iterations, in order to include the good properties of both moves. Our general recommendation is to consider more than one move type for updating branch lengths. However, our recommendation does not restrict to the BLM and BLNA moves. A sampler with good estimation performance could also be designed by including other move types, as long as irreducibility of the chain holds. An additional consideration is to check that under such moves, the chain is able to escape from computational trapping states.

The key result of Chapter 4 is an MCMC algorithm for estimating rates of substitution and stationary probabilities that achieves good estimation performance without the need to resort to computationally expensive tempered MCMC techniques. The Dirichlet proposal, which is the common choice for updating substitution rates and stationary probabilities in the phylogenetics literature, is shown to be unable to prevent the chain from falling into trapping states. The algorithm falls into a trap at the zero-boundary since the step-size of the proposal approaches zero as the current state tends to zero. This creates an ill-defined cycle in which the sampler keeps proposing candidate values very close to the current value because the step-size of the proposal is nearly zero. But the step-size is nearly zero because the current state is close to zero too and, therefore, proposed states are also nearly zero and the cycle starts again. The chain then shows many instances in its path in which it was unable to leave the zero-boundary for several iterations. This has detrimental effects on mixing. The solution that we proposed in Section 4.4.6 consists on shifting the centre of the Dirichlet proposal by a small quantity $\varepsilon > 0$. The ε-corrected algorithm is able to escape from trapping states at the zero-boundary and most importantly, at no extra computational cost. Chapter 6 showed that the ε-correction can also be effectively used in the proposal that generates candidate mixture proportions. When estimating the $Q + t$ mixture with a non-corrected proposal for mixture proportions, the algorithm becomes trapped when one of the mixture components is allocated very few observations. Once again, the ε-correction enables the algorithm to escape from trapping states while creating no extra computational burden.

Concerning parameter estimation, some applications of phylogenetic methods tend to report a 'most likely' tree, given the observed data, but pay less attention to other phylogenetic parameters. From our experience of working with biologists, it appears that the tree is usually regarded as the most interesting output of the analysis while less weight is given to substitution rates, stationary probabilities or branch lengths of the tree. Our classification treatment, based on heterogeneity that arises from the branch lengths and the Q-matrix parameters, is a call to phylogeneticists to redirect attention to features of the model that are perhaps not as easily interpretable as a tree, but that can also bring useful insights. For instance, our analysis of the *Borrelia burgdorferi* data, specifically the analysis of the *housekeeping gene* alignment with a two-component $Q + t$ mixture, yielded an estimated MAP tree with very low posterior support (Section 8.4.2). This suggests that several competing

trees are nearly as well supported by the data as the MAP tree itself. (A similar behaviour was observed in an equivalent analysis with the software package $\mathrm{MrBayes}$.) Conclusions that focus on this barely-supported posterior tree will miss other important aspects that can also be studied from the model. For instance, what is the rate at which transitions occur relative to transversions? Or, what is the total interior branch length and what does it tell us about the organisms' evolutionary processes? Within the context of heterogeneous data, do different sites share rules of nucleotide substitution, that is, do they obey the same Q-matrix? Or, do different sites agree in nucleotide composition? Questions like these need to be answered by the model parameters that represent these features, namely, the rates of nucleotide substitution and the stationary probabilities. In our opinion, it is important that phylogeneticists only give much credit to an estimated tree if that tree is well supported by the data. In applications where the observed data have approximately equal support for several competing trees, conclusions should not be limited to a single 'most likely' tree. Instead, reporting the posterior distribution for trees would be a fairer treatment of the uncertainty in the data. Alternatively, it seems more informative to choose common elements between the supported trees instead of picking a single 'most likely' one. This is the approach followed by methods that return a *consensus tree* (Sections 4.5 and 8.4.2). We believe that it is more adequate to summarise the tree-output via consensus trees than via MAP estimates, and this could be considered an area of further research. An extension like this to our method would be straightforward, as a consensus tree would be constructed from the MCMC output that our sampler already produces.

9.3 Analysis of DNA data

Chapters 7 and 8 analysed the DNA alignments of nine primates and sixteen strains of the bacterium *B. burgdorferi*, respectively. The purpose in Chapter 7 was to validate the proposed classification methodology by applying it to a well-understood dataset in the phylogenetics literature; the *primate mtDNA* alignment. An analysis of these data with a two-component $Q + t$ mixture model classified the sites according to their monomorphic or polymorphic nature. Sites that show no character variation at all, or monomorphic sites, clustered together into one evolutionary class, while sites that vary, or polymorphic, clustered in a different class. Traditional analyses assume that there are four evolutionary classes underlying these data, corresponding to the three codon positions and the tRNA region (e.g. [126, 63, 112]). The classification method that this thesis introduced did not detect evolutionary differences between the codon positions other than in terms of their monomorphicity/polymorphicity content. Codon positions that are freer to undergo substitution, such as $cp1$ or $cp3$, are generally polymorphic positions, while more conserved positions, such as $cp2$ or tRNA, mostly correspond to monomorphic sites. The grouping of the sites into a monomorphic and a polymorphic class creates a distinction between the codon positions in the sense that most $cp1$ and $cp3$ positions were allocated to the polymorphic class and most $cp2$ and tRNA were grouped into the monomorphic class. A further

analysis of only polymorphic sites (i.e. an alignment in which all monomorphic sites were removed) with a two-component $Q+t$ mixture, showed no clear segmentation according to codon positions. This suggests that the differences in rates of substitution between codon positions and tRNA reported by Yang [126], Larget and Simon [63] and Suchard, Weiss and Sinsheimer [112] may be caused by the effect that monomorphic sites have on each of these categories and not by evolutionary differences between the codon positions themselves. In other words, what our results indicated is that the codon positions themselves are not evolving differently but that monomorphic sites occur in different proportions at each of the codon positions, creating a distinction between $cp1$, $cp2$, $cp3$ and tRNA. It is monomorphic sites that 'evolve' differently to polymorphic ones regardless of the codon position from which they originate. Therefore, a partition into four classes (three for the codon positions plus a tRNA class) does not seem to be the most adequate one for the *primate mtDNA* alignment.

The analyses of the *Borrelia burgdorferi* data in Chapter 8 had the purpose of testing whether a set of nine different genes supported similar evolutionary processes or not. In this study, the true membership of a site to one or another gene was known. Evolutionary consistency between two genes was then suggested if sites originating from two distinct genes were classified to the same evolutionary class. In contrast, two genes were considered inconsistent in evolution if sites originating from them were grouped to different classes. The results from these analyses suggested evolutionary consistency between eight housekeeping genes and evolutionary inconsistency between these eight genes and a ninth gene, called the *ospC* gene. In this case, all monomorphic sites were removed from the analysed alignments to be able to detect evolutionary heterogeneity among the polymorphic positions. The clear evolutionary differences found between the eight housekeeping genes and the *ospC* gene can be further understood by their biological origins. The housekeeping genes are located in the main DNA-storage structure of *B. burgdorferi*, while the *ospC* gene is stored in extra pieces of DNA material that the bacterium carries. We then say that the housekeeping genes are chromosomal-located while the *ospC* gene is plasmid-located. In [71], Margos and colleagues mention possible inferential biases of phylogenetic studies based on concatenations of both chromosomal and plasmid-located genes (e.g. [8, 2]). Up to now and as far as we are aware, no study had formally demonstrated evolutionary incompatibility between these two gene categories. The immediate implications of the results presented in Chapter 8 are that phylogenetic inferences based on concatenated alignments of the form *housekeeping genes*|*ospC gene* are compromised. The conflicting evolutionary signals underlying such a concatenation generate spurious estimates that do not reflect the true nature of neither the chromosomal nor the plasmid-located classes. These results are based on the eight housekeeping genes chosen by Margos et al. [71] for their MLST scheme, and further research would be required to determine whether these conclusions hold more generally.

An interesting area of future research would be to apply the classification methodology to the original *ospC* alignment with both monomorphic and polymorphic sites included. The original *ospC* alignment has a reasonable proportion of monomorphic sites (around 51%) and *ospC* is a protein-coding gene. This means that an *ospC* sequence naturally groups into sets of three non-overlapping nucleotides to form codons, and different positions within a codon are believed to undergo substitutions at different rates (Section 3.3.3). The *ospC* alignment could be analysed in a similar way as the *primate mtDNA* data was studied (Chapter 7). If the classification procedure showed that sites group according to their monomorphic and polymorphic nature, this would support the results observed in Chapter 7. In other words, similar results to the ones obtained in Chapter 7 would corroborate that heterogeneity in protein-coding sequences is due to the evolutionary differences between monomorphic and polymorphic sites and not to the codon positions.

One of the most relevant features of our classification method is that it does not require us to rearrange the DNA alignment according to *a priori* known classes such as the change-point model by Suchard et al. [111] requires. It does not rely either on a prior specification of allocation variables in the form of a Markov chain (i.e. as a hidden Markov model) to capture the dependence between observations. The $Q + t$ mixture model captures the class-structure in the data without assuming that the allocation for a site depends on the underlying allocation of the previous site. This makes our classification method suitable for analysing a wide range of data, including sequences in which the true classes are not *a priori* known or alignments that have been rearranged in a way in which correlated sites are no longer close to one another. This thesis demonstrated the power and flexibility of phylogenetic mixture modelling as a means of accounting for heterogeneous data, identifying the classes that underlie those data and providing an interpretation of the nature of the classes. As computing power has increased phylogeneticists have been able to fit more realistic models without having to resort to as many possibly untenable assumptions about the processes that underlie the data. Phylogenetic mixture modelling is a step towards such extended realism.

Bibliography

[1] Akaike, H. New look at statistical-model identification. *IEEE Transactions of Automatic Control AC19*, 6 (1974), 716–723.

[2] Attie, O., Bruno, J. F., Xu, Y., Qiu, D., Luft, B. J., and Qiu, W. G. Co-evolution of the outer surface protein C gene (ospC) and intraspecific lineages of *Borrelia burgdorferi* sensu stricto in the northeastern United States. *Infection, Genetics and Evolution 7* (2007), 1–12.

[3] Behrens, G., Friel, N., and Hurn, M. Tuning tempered transitions. Unpublished.

[4] Binder, D. Bayesian cluster analysis. *Biometrika* (1978), 31–38.

[5] Brooks, S. P. Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society D 47*, 1 (1998), 69–100.

[6] Brown, E. W. Molecular differentiation of bacterial strains. In *Molecular Epidemiology*, M. Carrington and A. Rus Hoelzel, Eds. Oxford University Press, New York, USA, 2001, pp. 29–66.

[7] Brown, W. M., Prager, E. M., Wang, A., and Wilson, A. C. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution 18* (1982), 225–239.

[8] Bunikis, J., Garpmo, U., Tsao, J., Berglund, J., Fish, D., and Barbour, A. G. Sequence typing reveals extensive strain diversity of the Lyme borreliosis agents *Borrelia burgdorferi* in North America and *Borrelia afzelii* in Europe. *Microbiology 150* (2004), 1741–1755.

[9] Casjens, S., Palmer, N., van Vugt, R., Mun Huang, W., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R. J., Haft, D., Hickey, E., Gwinn, M., White, O., and Fraser, C. M. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Molecular Microbiology 35*, 3 (2000), 490–516.

[10] Cavalli-Sforza, L. L., and Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Evolution 21* (1967), 550–570.

[11] Celeux, G., Hurn, M., and Robert, C. P. Computational and inferencial difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*, 451 (2000), 957–970.

[12] Chatfield, C. *The analysis of time series. An introduction*, fifth ed. Chapman & Hall, Great Britain, 1996.

[13] Chen, D., Burleigh, J. G., and Fernández-Baca, D. Spectral Partitioning of Phylogenetic Data Sets Based on Compatibility. *Systematic Biology 56*, 4 (2007), 623–632.

[14] Chib, S., and Greenberg, E. Understanding the Metropolis-Hastings Algorithm. *The American Statistician 49*, 4 (1995), 327–335.

[15] Chib, S., and Jeliazkov, I. Marginal Likelihood from the Metropolis-Hastings Output. *Journal of the American Statistical Association 96*, 453 (2001), 270–281.

[16] Crozier, R. H., and Crozier, Y. C. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics 133* (1993), 97–117.

[17] Diaconis, P. W., and Holmes, S. P. Matchings and phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America 95*, 25 (DEC 8 1998), 14600–14602.

[18] Diebolt, J., and Robert, C. P. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B 56* (1994), 363–375.

[19] Edwards, A. W. F., and Cavalli-Sforza, L. L. Reconstruction of evolutionary trees. In *Phenetic and Phylogenetic Classification*, V. H. Heywood and J. McNeill, Eds., no. 6. Systematics Association Publ., London, 1964, pp. 67–76.

[20] Farris, J. S. Estimating phylogenetic trees from distance matrices. *The American Naturalist 106*, 951 (1972), 645–668.

[21] Felsenstein, J. The Newick tree format. Department of Genome Sciences, University of Washington , Seattle, USA. Available at: http://evolution.gs.washington.edu/phylip/newicktree.html [Accessed 18 May 2009].

[22] Felsenstein, J. Maximum likelihood and minimum-step methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology 22* (1973), 240–249.

[23] Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17* (1981), 368–376.

[24] Felsenstein, J. Statistical Inference of Phylogenies. *Journal of the Royal Statistical Society A 146*, 3 (1983), 246–272.

[25] Felsenstein, J. PHYLIP: Phylogeny Inference Package (Version 3.2). *Cladistics 5* (1989), 164–166.

[26] Felsenstein, J. The troubled growth of statistical phylogenetics. *Systematic Biology 50*, 4 (2001), 465–467.

[27] Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates Inc., USA, 2004.

[28] Felsenstein, J., and Churchill, G. A. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution 13*, 1 (1996), 93–104.

[29] Fitch, W. M., and Margoliash, E. Construction of phylogenetic trees. *Science 155* (1967), 279–284.

[30] Fitch, W. M., and Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics 4* (1970), 579–593.

[31] Gamerman, D., and Lopes, H. F. *Markov chain Monte Carlo. Stochastic Simulation for Bayesian Inference*, second ed. Texts in Statistical Science. Chapman & Hall/CRC, USA, 2006.

[32] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, second ed. Chapman & Hall, USA, 2004.

[33] Geman, S., and Geman, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 6 (1984), 721–741.

[34] Gesell, T., and von Haeseler, A. In silico sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics 22*, 6 (2005), 716–722.

[35] Geyer, C. J. Markov-chain Monte-Carlo Maximum-Likelihood. In *Computing Science and Statistics* (1991), Keramidas, E. M., Ed., pp. 156–163.

[36] Geyer, C. J. Practical Markov Chain Monte Carlo. *Statistical Science 7*, 4 (1992), 473–511.

[37] Geyer, C. J., and Thompson, E. A. Annealing Markov-chain Monte-Carlo with applications to ancestral inference. *Journal of the American Statistical Association 90*, 431 (1995), 909–920.

[38] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. Chapman & Hall, Great Britain, 1996, pp. 1–19.

[39] Gilks, W. R., and Roberts, G. O. Strategies for improving MCMC. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. Chapman & Hall, Great Britain, 1996, pp. 89–114.

[40] Golding, G. B. Estimates of DNA and protetin sequence divergence: an examination of some assumptions. *Molecular Biology and Evolution 1*, 1 (1983), 125–142.

[41] Green, P. J. Discussion on representations of knowledge in complex systems (by U. Grenander and M. I. Miller). *Journal of the Royal Statistical Society B 56* (1994), 589–590.

[42] Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82* (1995), 711–732.

[43] Green, P. J., and Han, X.-L. Metropolis methods, Gaussian proposals, and antithetic variables. *Lecture Notes in Statistics 74* (1992), 142–164.

[44] Grimmett, G. R., and Stirzaker, D. R. *Probability and Random Processes*, third ed. Oxford University Press, Great Britain, 2004.

[45] Hasegawa, M., Kishino, H., and Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution 22* (1985), 160–174.

[46] Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 1 (1970), 97–109.

[47] Haubold, B., and Wiehe, T. *Introduction to Computational Biology. An Evolutionary Approach*, first ed. Birkhäuser Verlag, Germany, 2006.

[48] Hayasaka, K., Gojobori, T., and Horai, S. Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution 5*, 6 (1988), 626–644.

[49] Huber, K. T., and Moulton, V. Phylogenetic networks. In *Mathematics of Evolution and Phylogeny*, O. Gascuel, Ed., first ed. Oxford University Press, Great Britain, 2005, pp. 178–204.

[50] Huelsenbeck, J. P., Larget, B., and Alfaro, M. E. Bayesian Phylogenetic Model Selection using Reversible Jump Markov chain Monte Carlo. *Molecular Biology and Evolution 21*, 6 (2004), 1123–1133.

[51] Huelsenbeck, J. P., and Ronquist, F. MrBayes: Bayesian inference of phylogeny. *Bioinformatics 17* (2001), 754–755.

[52] Husmeier, D. Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics 21*, Suppl. 2 (2005), 166–172.

[53] Husmeier, D., and McGuire, G. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. *Molecular Biology and Evolution 20*, 3 (MAR 2003), 315–337.

[54] Husmeier, D., and Wright, F. Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology 8*, 4 (2001), 401–427.

[55] Izenman, A. J., and Sommer, C. J. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association 83*, 404 (1988), 941–953.

[56] Jasra, A., Holmes, C. C., and Stephens, D. A. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science 20*, 1 (2005), 50–67.

[57] Jukes, T., and Cantor, C. Evolution of Protein Molecules. In *Mammalian Protein Metabolism*, H. N. Munro, Ed. Academic Press, New York, USA, 1969, pp. 21–132.

[58] Kass, R. E., and Raftery, A. E. Bayes Factors. *Journal of the American Statistical Association 90*, 430 (1995), 773–795.

[59] Kimura, M. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution 16* (1980), 111–120.

[60] King, J. L., and Jukes, T. H. Non-Darwinian Evolution. *Science 164* (1969), 788–798.

[61] Kolaczkowski, B., and Thornton, J. W. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution 25*, 6 (JUN 2008), 1054–1066.

[62] Lakner, C., van der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Systematic Biology 57*, 1 (2008), 86–103.

[63] Larget, B., and Simon, D. L. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution 16*, 6 (1999), 750–759.

[64] Lartillot, N., Brinkmann, H., and Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology 7*, (Suppl. 1)S4 (2007).

[65] Li, S., Pearl, D. K., and Doss, H. Phylogenetic tree construction using Markov Chain Monte Carlo. *Journal of the American Statistical Association 95* (2000), 493–508.

[66] Li, W. H., Tanimura, M., and Sharp, P. M. Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution 5* (1988), 313–330.

[67] Lockhart, P., Novis, P., Milligan, B. G., Riden, J., Rambaut, A., and Larkum, T. Heterotachy and tree building: A case study with plastids and eubacteria. *Molecular Biology and Evolution 23*, 1 (JAN 2006), 40–45.

[68] Lockhart, P. J., Larkum, A. W. D., Steel, M. A., Waddell, P. J., and Penny, D. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America 93*, 5 (MAR 5 1996), 1930–1934.

[69] Lopez, P., Casane, D., and Philippe, H. Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution 19*, 1 (JAN 2002), 1–7.

[70] Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences USA 95* (1998), 3140–3145.

[71] Margos, G., Gatewood, A. G., Aanensen, D. M., Hanincova, K., Terekhova, D., Vollmer, S. A., Cornet, M., Piesman, J., Donaghy, M., Bormane, A., Hurn, M. A., Feil, E. J., Fish, D., Casjens, S., Wormser, G. P., Schwartze, I., and Kurtenbach, K. MLST of housekeeping genes captures geographic population structure and suggests a European origin of *Borrelia burgdorferi*. *Proceedings of the National Academy of Sciences USA 105*, 25 (2008), 8730–8735.

[72] Marinari, E., and Parisi, G. Simulated tempering. A new Monte-Carlo scheme. *Europhysics Letters 19*, 6 (1992), 451–458.

[73] Marttinen, P., Baldwin, A., Hanage, W. P., Dowson, C., Mahenthiralingam, E., and Corander, J. Bayesian modeling of recombination events in bacterial populations. *BMC Bioinformatics 9* (2008).

[74] Maslow, J. N., Mulligan, M. E., and Arbeit, R. D. Molecular Epidemiology: Application of Contemporary Techniques to the Typing of Microorganisms. *Clinical Infectious Diseases 17* (1993), 153–164.

[75] Mau, B., and Newton, M. A. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics 6* (1997), 122–131.

[76] McLachlan, G., and Peel, D. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, USA, 2000.

[77] Meade, A., and Pagel, M. A Phylogenetic Mixture Model for Heterotachy. In *Evolutionary Biology from Concept to Application*, P. Pontarotti, Ed., first ed. Springer-Verlag, 2008, pp. 29–41.

[78] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*, 6 (1953).

[79] Minin, V. N., Dorman, K. S., Fang, F., and Suchard, M. A. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics 21* (2005), 3034–3042.

[80] Modica-Napolitano, J. S., and Singh, K. K. Mitochondria as targets for detection and treatment of cancer. *Expert Reviews in Molecular Medicine 4* (2002), 1–19.

[81] Moore, G. W., Goodman, M., and Barnabas, J. An iterative approach from the stand-point of the additive hypothesis to the dendrogram problem posed by molecular data sets. *Journal of Theoretical Biology 38* (1973), 423–457.

[82] Morgan, B. J. T. *Elements of Simulation*, first ed. Chapman & Hall, Great Britain, 1984.

[83] Moritz, C., and Hillis, D. M. Molecular Systematics: Context and Controversies. In *Molecular Systematics*, D. M. Hillis, C. Moritz, and B. K. Mable, Eds., second ed. Sinauer Associates, Inc., Sunderland, MA, USA, 1996, pp. 1–13.

[84] Neal, R. M. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing 6*, 4 (1996), 353–366.

[85] Neyman, J. Molecular studies of evolution: A source of novel statistical problems. In *Statistical Decision Theory and Related Topics*, S. S. Gupta and J. Yackel, Eds. Academic Press, New York, 1971, pp. 1–27.

[86] Ochman, H., Lawrence, J., and Groisman, E. Lateral gene transfer and the nature of bacterial innovation. *Nature 405*, 6784 (2000), 299–304.

[87] Page, R. D. M., and Holmes, E. C. *Molecular Evolution. A phylogenetic approach*. Blackwell Publishing, Great Britain, 1998.

[88] Pagel, M., and Meade, A. A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data. *Systematic Biology 56*, 4 (2004), 571–581.

[89] Pagel, M., and Meade, A. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philosophical Transactions of the Royal Society B 363* (2008), 3955–3964.

[90] Philippe, H., Casane, D., Gribaldo, S., Lopez, P., and Meunier, J. Heterotachy and functional shift in protein evolution. *IUBMB Life 55*, 4-5 (APR-MAY 2003), 257–265.

[91] Raftery, A. E. Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. Chapman & Hall, Great Britain, 1996, pp. 163–187.

[92] Rambaut, A., and Grassly, N. C. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences 13* (1997), 235–238.

[93] Rannala, B., and Yang, Z. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution 43* (1996), 304–311.

[94] Richardson, S., and Green, P. J. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B 59*, 4 (1997), 731–758.

[95] Ripley, B. D. *Stochastic Simulation*. John Wiley & Sons, New York, USA, 1987.

[96] Robert, C. P. Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. Chapman & Hall, Great Britain, 1996, pp. 441–464.

[97] Roberts, G. O. Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. Chapman & Hall, Great Britain, 1996, pp. 45–57.

[98] Robinson, D. F. Comparison of Labeled Trees with Valency Three. *Journal of Combinatorial Theory 11* (1971), 105–119.

[99] Rocha, E. P. C., Touchon, M., and Feil, E. J. Similar compositional biases are caused by very different mutational effects. *Genome Research 16*, 12 (2006), 1537–1547.

[100] Rodriguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B., and Melkonian, M. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta. *Molecular Biology and Evolution 24*, 3 (MAR 2007), 723–731.

[101] Roeder, K. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association 85*, 411 (SEP 1990), 617–624.

[102] Ronquist, F., and Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19* (2003), 1572–1574.

[103] Ronquist, F., Huelsenbeck, J. P., and van der Mark, P. MrBayes 3.1 Manual. Available at: http://mrbayes.csit.fsu.edu/manual.php [Accessed 18 May 2009].

[104] Schwarz, G. Estimating dimension of a model. *Annals of Statistics 6*, 2 (1978), 461–464.

[105] Semple, C., and Steel, M. *Phylogenetics*. Oxford University Press, USA, 2003.

[106] Shores, T. *Applied Linear Algebra and Matrix Analysis*. Springer New York, 2007.

[107] Silverman, B. W. *Density estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[108] Simon, D. L., and Larget, B. BAMBE: Bayesian analysis in molecular biology and evolution [online]. Department of Mathematics and Computer Science, Duquesne University, Pennsylvania, USA, 1998. Available at: http://www.mathcs.duq.edu/larget/bambe.html [Accessed 16 May 2009].

[109] Singh, K. K. *Mitochondrial DNA mutation in aging, disease and cancer*. Springer, New York, USA, 1998.

[110] Stanek, G., Strle, F., Gray, J., and Wormser, G. P. History and Characteristics of Lyme Borreliosis. In *Lyme Borreliosis. Biology, Epidemiology and Control*, J. S. Gray, O. Kahl, R. S. Lane, and G. Stanek, Eds. CABI Publishing, 2002, pp. 1–28.

[111] Suchard, M. A., Weiss, R. E., Dorman, K. S., and Sinsheimer, J. S. Inferring spatial phylogenetic variation along nucleotide sequences: a multiple changepoint model. *Journal of the American Statistical Association 98*, 462 (2003), 427–437.

[112] Suchard, M. A., Weiss, R. E., and Sinsheimer, J. S. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution 18*, 6 (2001), 1001–1013.

[113] Swofford, D. L. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta. Computer program distributed by Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, 2002.

[114] Tamura, K., Dudley, J., Nei, M., and Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution 24*, 8 (2007), 1596–1599.

[115] Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. In *Lectures on Mathematics in the Life Sciences*, R. M. Miura, Ed., vol. 17. American Mathematical Society, Providence, USA, 1986, pp. 57–86.

[116] Titterington, D. M., Smith, A. F. M., and Makov, U. E. *Statistical Analysis of Finite Mixture Distributions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Great Britain, 1985.

[117] Uzzell, T., and Corbin, K. W. Fitting Discrete Probability Distributions to Evolutionary Events. *Science 172*, 3988 (1971), 1089–1096.

[118] van Belkum, A., Struelens, M., de Visser, A., Verbrugh, H., and Tibayrenc, M. Role of Genomic Typing in Taxonomy, Evolutionary Genetics, and Microbial Epidemiology. *Clinical Microbiology Reviews 14*, 3 (2001), 547–560.

[119] Wakeley, J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology and Evolution 11*, 4 (1996), 158–163.

[120] Watson, J. D., and Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature 171*, 4356 (1953), 737–738.

[121] Webb, A., Hancock, J. M., and Holmes, C. C. Phylogenetic inference under recombination using Bayesian stochastic topology selection. *Bioinformatics 25*, 2 (2009), 197–203.

[122] Wolfe, J. H. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research 5* (1970), 329–350.

[123] Yang, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution 10*, 6 (1993), 1396–1401.

[124] Yang, Z. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution 39* (1994), 105–111.

[125] Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution 39* (1994), 306–314.

[126] Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics 139* (1995), 993–1005.

[127] Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* (1997), 555–556.

[128] Yang, Z. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, Great Britain, 2006.

[129] Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution 24* (2007), 1586–1591.

[130] Yang, Z., Goldman, N., and Friday, A. Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation. *Molecular Biology and Evolution 11*, 2 (1994), 316–324.

[131] Yang, Z., and Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution 14*, 7 (1997), 717–724.